# Journal of the American Society for Information Science

# JASIS



The *Journal of the American Society for Information Science* (ISSN: 0002-8231) is the official journal of the American Society for Information Science, **8720 Georgia Avenue, Suite 501, Silver Spring, MD 20910; (301) 495-0900, asis@cni.org.** It is published monthly, except twice in April and May for the Society by John Wiley & Sons, Inc. The positions expressed by the contributors do not necessarily reflect the position or the official policy of the Society.



**Subscription Price** (Volume 49, 1998): $999.00 in US, $1139.00 in Canada and Mexico, $1244.00 outside North America. All subscriptions outside US will be sent by air. Payment must be made in US dollars drawn on a US bank. Claims for undelivered copies will be accepted only after the following issue has been received. Please enclose a copy of the mailing label. Missing copies will be supplied when losses have been sustained in transit and where reserve stock permits. Please allow four weeks for processing a change of address. For subscription inquiries, please call (212) 850-6645; E-mail: SUBINFO@wiley.com

**Postmaster:** Send address changes to The *Journal of the American Society for Information Science.* Caroline Rothaug, Director, Subscription Fulfillment and Distribution. John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158. For subscription inquiries, please call customer service at (212) 850-6645 or write to the above address.



**Other Correspondence:** Address all other correspondence to: Publisher, The *Journal of the American Society for Information Science,* Professional, Reference, and Trade Group, John Wiley & Sons, Inc., 605 Third Avenue, New York, NY 10158.

The contents of this journal are indexed in the following: *Chemical Abstracts, CompuMath, Computer Literature Index, Current Contents/Engineering Technology and Applied Sciences, Current Contents/Social and Behavioral Sciences, Current Index to Journals in Education (CIJE), Engineering Index, Information Science Abstracts, INIST for the PASCAL database, ISI Compumath, Library and Information Science Abstracts (LISA), Library Literature, Research Alert (ISI),* and *Social Sciences Citation Index.*

**Editorial Production, John Wiley:** Lillian Solondz

**Cover Design, John Wiley:** Adrienne Weiss

(∞) This paper meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

Journal of the American Society for Information Science

# JASIS

VOLUME 49 NUMBER 13 NOVEMBER 1998

## CONTENTS







Volume 49, Number 13 was mailed the week of October 12, 1998

# In This Issue

We begin with two articles suggesting the possible separation of document and query vector space. Viewing information retrieval as a topology on a document space determined by a similarity function between queries and documents gives what Egghe and Rousseau call a retrieval topology. Such topologies might use a pseudo metric which measures the distance between documents independent of the query space, or might make all similarity functions between documents and queries continuous, called here the similarity topology. The topological model allows the introduction of Boolean operators. The inner product is suggested as producing a more powerful model than the cosine measure. Bollmann-Sdorra and Raghavan show that if query term weights are to be useful in retrieval, term independence is an undesirable property in a query space. Independence remains desirable in document space. It would appear that the assumptions that documents and queries are elements of the same space, and that term independence is required, are not warranted.

Al-Fares presents a new loan policy model which incorporates a decision variable for maximum books to be borrowed, along with the traditional loan period, and adds user satisfaction with policies to the usual book availability satisfaction indicator. Each indicator is defined as the ratio of satisfied demand to total demand. Number of renewals, duplications, demand, and reservations are considered to have a very small effect.

Yager and Rybalov assume $m$ retrieval systems without file overlap each providing a ranked list of texts based upon their varying ranking criteria, and in response to a common query, and define fusion as the construction of a single ordered list of the $n$ most relevant texts over all $m$ system responses. This requires determining the potential of each system to provide relevant answers to the query.

A previous fusion method which is empirically effective but where different runs will result in different orderings, uses a random selection method biased toward the length of the contributing list. Alternatively one might use the longest remaining list for each choice or take equally from each collection until the shortest is exhausted, and then continue until the next shortest is exhausted, and on, until all are exhausted. A centralized fusion scheme computes a value based upon the number of documents in a list and the number already removed. The value is re-computed for each collection after each removal of the top element in the collection with the highest value. Another possibility is a proportional approach, where the list value is its remaining number of elements less one divided by the original number, and a value can be assigned to each individual document which is the number of elements in the list less its position in the list, divided by the number of elements in the list.

Bates provides a review of what we know and do not know about indexing and access that will apply to large digital document files. Particularly she emphasizes that statistical regularities exist in the subject representation of files and should influence design, that subject domain should affect system design, and that what we know of human linguistic and searching behavior must be taken into account for an optimal information retrieval system.

The indexing for 16,691 documents from 1982 to 1994 which were assigned at least one term from the software engineering category was collected by Coulter, Monarch, and Konda and a co-occurrence study carried out to determine the interaction of software engineering areas of study over time. The association measure was the square of the co-occurrences of two terms over the product of their occurrences. The threshold value was varied with the size of the data sets, but the number of links and nodes was fixed at twenty-four and twenty. For the period 1982–1986 15 networks were generated; for 1987–1990 16; and for 1991–1994 11. The networks exhibit considerable change over time although some consistent themes, like software development and user interfaces, persist.

To address the role of information technology in non-traditional organizations Travica treats IT as level of use of several specific technologies, and non-traditional structure as the level of organization structure, plus other selected variables. Data came from surveys of a random stratified sample of employees at twelve local accounting offices and an interview with the local manager. Information technology correlates with non-traditional structure. Information technology correlates negatively with formalization and centralization, and positively with cross boundary communication. Spatial dispersion is negatively associated with trust sharing.

**Bert R. Boyce**
*Louisiana State University*

# Topological Aspects of Information Retrieval

**Leo Egghe,**
*LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium and UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium. E-mail: legghe@luc.ac.be*

**Ronald Rousseau**
*UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium and KHBO, Zeedijk 101, B-8400 Oostende, Belgium. E-mail: rousseau@kh.khbo.be*

**Let (*DS, DQ,* sim) be a retrieval system consisting of a document space *DS,* a query space *QS,* and a function sim, expressing the similarity between a document and a query. Following D. M. Everett and S. C. Cater (1992), we introduce topologies on the document space. These topologies are generated by the similarity function sim and the query space *QS.* Three topologies will be studied: The retrieval topology, the similarity topology, and the (pseudo-)metric one. It is shown that the retrieval topology is the coarsest of the three, while the (pseudo-) metric is the strongest. These three topologies are generally different, reflecting distinct topological aspects of information retrieval. We present necessary and sufficient conditions for these topological aspects to be equal. Several examples of topological retrieval systems are presented. One of these examples is a vector space model that yields a simplification of the Everett-Cater model, yet having a more diversified spectrum of topological properties. Finally, it is shown that information retrieval based on Boolean operators is an intrinsic part of the general topological model. This is a major motivation of the introduction of topologies in theoretical IR models.**

## 1. Introduction

Theoretical document retrieval has two main components: indexing and searching. As our article only deals with theoretical aspects, we leave all software and hardware aspects aside. Indexing determines the way documents are placed in a database. We will use the term "indexing" in a very broad way: Even when the original document has not changed, e.g., it is directly put on the Internet, we call this a form of indexing. Therefore, indexing is considered as a surjective map, $f$, from the set of original documents, $X$, to the set of "indexed document representation," denoted as $DS$, the document space:

$$f: X \rightarrow DS. \tag{1}$$

Here "surjective" means that every element in $DS$ is the representation of at least one document in $X$. Elements of $DS$ will be denoted by $D$, $E$, $D'$, etc. They can be bibliographical records with or without abstracts, or full-text documents, or any form of multimedia document.

Searching uses queries and refers to the way customers express an information need, a topic on which they want to be informed. Users' needs are also transformed, and are represented by a formal query. The set of all formal queries is called query space and denoted by $QS$. So, again we can consider a surjective map, $g$, from the space of all topics (or use needs) to the query space:

$$g: Y \rightarrow QS. \tag{2}$$

From now on elements in $QS$ will be denoted by the symbols $Q$, $Q'$, etc. We will mainly be interested in single attribute queries, e.g., keywords or authors. Different query definitions, hence different $QS$ spaces lead to different retrieval systems.

The spaces $DS$ and $QS$ do not completely determine a retrieval system. What is missing is a way to express the degree of agreement between a query $Q \in QS$ and a document $D \in DS$. This is done using a similarity function, denoted as sim:

$$\text{sim}: DS \times QS \rightarrow \mathbf{R} : (D, Q) \rightarrow \text{sim}(D, Q).$$

We will often keep $Q$ fixed, which yields functions of the form:

$$\text{sim}(., Q): DS \rightarrow \mathbf{R} : D \rightarrow \text{sim}(D, Q). \tag{3}$$

Received December 19, 1996; revised October 23, 1997; accepted October 23, 1997.

The higher the value of $\text{sim}(D, Q)$, the more $D$ and $Q$ correspond and therefore, the more probable it will be that the document $D$ satisfies the user's need. Consequently, such documents are wanted in the set of retrieved documents. This retrieved documents set is further determined by a cutoff value $r \in \mathbf{R}$. So, all documents $D \in DS$, such that

$$\text{sim}(D, Q) > r \tag{4}$$

are retrieved.

Other forms of thresholds such as

$$r_1 < \text{sim}(D, Q) < r_2 \tag{5}$$

or even clustering techniques can also be used (Salton & McGill, 1983).

Everett and Cater (1992) have shown that requirements 4 and 5 can be used to determine topological properties of the space $DS$. The definition of a topology and of a topological space is given in Appendix A. As such, it is not clear that such systems express "how near" points (here documents) are to each other. However, many topologies are determined by functions that measure similarity or distance between two documents (see Equations 4, 5, or the example of a metric space as a special case of a topological space—Appendix A) and, hence, the link with this powerful mathematical tool is clear.

Topologies on $DS$ as determined by Equations 4 or 5, or, in general, by using queries $Q \in QS$, are called topological retrieval systems. Indeed, the above described available notion of "how near" documents are, can be used in IR. It is a kind of predetermined system of documents for which the notion of "relatedness" is available and this is independent from a used query $Q$ (i.e., prior to the retrieval action itself, but with the knowledge of $QS$, the set of all possible queries).

Different topologies will, in addition, yield different retrieval properties. As such, the notion of a topological retrieval system is closely linked to the dynamics of retrieval behavior. The mathematical tools made available through these systems allow us to give answers to questions such as: "How do slight alterations in queries or thresholds influence the retrieval result?"

In the next section, we will set up a general framework for topological retrieval. Three topologies on $DS$ will play a role: the retrieval topology $\tau$, the (pseudo-)metric topology $\tau'$ (both introduced by Everett & Cater, 1992), and the similarity topology $\tau''$.

In the second section, we present examples of document spaces $DS$ for which the three topologies coincide, as well as examples where any two of these topologies are different. These properties illustrate the specific role each of these topologies plays. In particular, we give an example of a vector space model on $DS = I^n$ ($I$ denotes the unit interval $[0, 1]$, $n \in N$) with the inner product as similarity function. We will show that this approach leads to a simpler model, with more interesting properties, than the vector model studied in Everett and Cater (1992), which is based on Salton's cosine measure.

The last section shows that information retrieval based on Boolean connectives AND and OR is an intrinsic part of the general topological model. In particular, it is only necessary to have a query space $QS$ consisting of elementary queries. The Boolean aspects are taken care of by the topology of $DS$. This generalizes the Boolean algebra results of Cater (1986) and Everett and Cater (1992).

Mathematical proofs of most of the results are relegated to the Appendices.

**Notation.** We will use the abbreviation iff for the phrase "if and only if."

## 2. Document Spaces as Topological Spaces

A retrieval system is a triple ($DS$, $QS$, sim) consisting of a document space $DS$, a query space $QS$ and a similarity function sim. At this moment, $DS$ is only a set. However, using the query set $QS$ and the similarity function sim will enable us to put an additional structure on $DS$, namely that of a topological space. We focus here on the document space as a topological space (by using $QS$) since we want to have a notion of similarity between documents. But theoretically, there is no real difference between $QS$ and $DS$: We could study a topological system on $QS$, determined by $DS$. This is an example of a dual situation. For more on this, see Egghe and Rousseau (1997a).

At this point, we refer the reader who is not familiar with topological notions again to Appendix A, and the second section of Everett and Cater (1992). Information on general topology can be found in any of the following books: Császár (1978), Dugundji (1966), and Willard (1970). Special examples of topological spaces can be found in Steen and Seebach (1978). Finally, Wilansky (1970) and Kreyszig (1978) are good references on normed spaces.

In this article we will study three topologies on $DS$.

### *The Retrieval Topology $\tau$ (Everett & Cater, 1992)*

The retrieval topology, denoted as $\tau$, is generated by the subbasis

$$\{R(Q, r) | r \in \mathbf{R}, Q \in QS\} \tag{6}$$

where a retrieval $R(Q, r)$ is defined as:

$$R(Q, r) = \{D \in DS | \text{sim}(D, Q) > r\}. \tag{7}$$

A subbasis is a subset $C$ of the set of all open sets such that all open sets can be constructed by forming finite intersections of elements of $C$.

Note that, usually, the range of all sim(., $Q$) is in $\mathbf{R}^+$ (the positive real numbers, including zero), so that in these cases $\tau$ is generated by

$$\{R(Q, r) | r \in \mathbf{R}^+, Q \in QS\}.$$

There is no mathematical reason to limit the range of sim to the positive numbers. Therefore, we decided to use $\mathbf{R}$ in theoretical arguments, but to use $\mathbf{R}^+$ in all examples. The set

$$\text{ret}_\tau(Q) = \{R(Q, r) | r \in \mathbf{R}\} \tag{8}$$

is called the set of retrievals of $Q$ (or, in short, the retrievals of $Q$) w.r.t. $\tau$. Also, the sets $R(Q, r)$, $r \in \mathbf{R}$, $Q \in QS$ are called retrievals of the system ($DS$, $QS$, sim, $\tau$) or, in short, of $\tau$.

Note that the set of retrievals of a query $Q$ is the set of all possible answers to the query $Q$ in the system ($DS$, $QS$, sim) with the retrieval topology. Different answers are obtained by changing the threshold. Consequently, the retrievals of a fixed query $Q$, form a nested set, i.e.,

$$R(Q, r_1) \subset R(Q, r_2) \text{ iff } r_1 \geq r_2.$$

Another name for $\tau$ could be the "threshold" topology, a name that is clear from the definition 7. Here, a document is retrieved if its similarity with a query $Q$ is at least a a certain value $r$.

*The (Pseudo-)Metric Topology $\tau'$ (Everett & Cater, 1992)*

Everett and Cater (1992) give the following definition. Let ($DS$, $QS$, sim) be a retrieval system. Define a function $d$ on $DS \times DS$ as follows:

$$d(D, E) = \sup_{Q \in QS} |\text{sim}(D, Q) - \text{sim}(E, Q)|. \tag{9}$$

The function $d$ should be a pseudo-metric (see Appendix A for the definition), leading to the pseudo-metric topology $\tau'$. However, there is a problem with definition 9. It is possible that the supremum is infinite, in which case $d$ does not exist in the usual sense. To correct this, we will define $\tau'$ in a more accurate way. For the ease of the notation, we first define

$$\rho_Q(D, E) = |\text{sim}(D, Q) - \text{sim}(E, Q)| \tag{10}$$

here $D, E \in DS$, $Q \in QS$. Then define

(i) if the supremum of Equation 10 is finite:

$$d(D, E) = \sup_{Q \in QS} \rho_Q(D, E)$$

i.e., the pseudo-metric defined by Everett and Cater;

(ii) $d' : DS \times DS \to [0, 1]$ by

$$d'(D, E) = \sup_{Q \in QS} \frac{\rho_Q(D, E)}{1 + \rho_Q(D, E)} \tag{11}$$

(iii) $d'' : DS \times DS \to [0, 1]$ by

$$d''(D, E) = \sup_{Q \in QS} [\min(1, \rho_Q(D, E))]. \tag{12}$$

Clearly $d'$ and $d'$ always exist and are finite. The following theorem states that from a topological point of view, the pseudo-metrics $d$, $d'$, and $d''$ are "the same" (although they are, of course, different as metrics).

*Theorem 2.1*

Let $d$, $d'$, and $d''$ be as defined in Equations 9, 11, and 12. Then

(i) the (pseudo-)metrics $d'$ and $d''$ are equivalent
(ii) if $d$ exists, it is equivalent with $d'$ and $d''$.

The proof of this theorem is given in Appendix B. The topology generated by $d$ (if it exists), $d'$ or $d''$ is called the pseudo-metric topology on $DS$ and is denoted by $\tau'$. The topology $\tau'$ measures absolute distances between documents, independent of a used query in IR.

As noted by Everett and Cater (1992) the following result is true:

*Proposition 2.2*

$\tau'$ is generated by the subbasis

$$V(D, \varepsilon) = \{E \in DS | \rho_Q(D, E) < \varepsilon, \ \forall Q \in QS\}$$

where $\varepsilon$ is any strictly positive real number. The proof follows readily from the proof of Theorem 2.1.

*Definition 2.3 (Everett & Cater, 1992)*

The retrieval system ($DS$, $QS$, sim) separates the points of $DS$ if,

$$(\forall Q \in QS : \text{sim}(D, Q) = \text{sim}(E, Q)) \Rightarrow (D = E).$$

It is obvious that the pseudo-metrics above are metrics iff the retrieval system ($DS$, $QS$, sim) separates the points of $DS$.

*The Similarity Topology $\tau''$*

In order to obtain consistent results, stability is very important in IR. It is, e.g., necessary that if documents $D$ and $E$ are "alike" (an expression to be defined in each concrete case), their similarity values must also be close to

each other. In this way slight changes in, e.g., recall requirements, yield only slight changes in the retrieval result. Such a behavior can only be guaranteed if the functions sim(., $Q$) are continuous. We therefore define the topology $\tau''$, referred to as the similarity topology, as the coarsest topology on $DS$ that makes all similarity functions sim(., $Q$) continuous. The next theorem describes a subbasis of the similarity topology.

*Theorem 2.4*

The similarity topology $\tau''$ is generated by the subbasis

$$\{U(Q, r_1, r_2) | Q \in QS, r_1 < r_2\}$$

where

$$U(Q, r_1, r_2) = \{D \in DS | r_1 < \mathrm{sim}(D, Q) < r_2\}$$

$$= \mathrm{sim}(., Q)^{-1}(]r_1, r_2[)$$

with $r_1 < r_2$, $r_1, r_2 \in \mathbf{R}$, $Q \in QS$.

*Proof.* The proof is easy: It follows immediately from properties of the inverse relation and the fact that the open intervals $]r_1, r_2[$ are a basis for the Euclidean topology on the real line. □

The set

$$\mathrm{ret}_{\tau''}(Q) = \{U(Q, r_1, r_2) | r_1 < r_2\}$$

is called the set of retrievals of $Q$ (or, in short, the retrievals of $Q$) w.r.t. $\tau''$. Also, the sets $U(Q, r_1, r_2)$, $r_1 < r_2$, $r_1, r_2 \in \mathbf{R}$, $Q \in QS$, are called retrievals of the system ($DS$, $QS$, sim, $\tau''$) or, in short, of $\tau''$.

Note that if $r_3 < r_1 < r_2 < r_4$ then

$$U(Q, r_1, r_2) \subset U(Q, r_3, r_4).$$

The topology $\tau''$ describes the retrieval of documents according to their closeness to a given query $Q$. Here (in contrast with $\tau$) the exact query $Q$ must be matched. When it is clear from the context with which topology we work, we will drop the subscript $\tau$ or $\tau''$ and simply write ret($Q$).

Results on these topologies will follow in the next sections. We can, however, already point out one intrinsic property of these topological spaces. The topologies $\tau$, $\tau'$, and $\tau''$ determine neighborhoods around every document $D \in DS$. These neighborhoods determine a filtering of documents: The finer we look, the more closely we end up in the neighborhood of $D$. The availability of such neighborhoods determines the degree of "fine tuning" that is possible in IR-systems that work with $\tau$, $\tau'$, or $\tau''$. Otherwise stated, the topologies on the document space $DS$ determine a pre-defined structure on $DS$ of "what documents are close to (a) certain document(s)," even without the formulation of a specific query $Q \in QS$. It is the totality $QS$ of all possible queries that determines this pre-defined structure. This, in turn, is comparable (but totally different in nature) with the statistical clustering techniques for documents that exist—see, e.g., Salton and McGill (1983).

The three topologies we have introduced are related in the following way.

*Theorem 2.5*

$$\tau \subset \tau'' \subset \tau'$$

*Proof.* The sets $R(Q, r) = \mathrm{sim}(., Q)^{-1}(]r, +\infty[)$, generating the retrieval topology, are open in $\tau''$ since sim(., $Q$) is continuous on ($DS$, $\tau''$). Hence, $\tau \subset \tau''$. Furthermore, it is clear that sim(., $Q$) is continuous on ($DS$, $\tau'$), hence $\tau'' \subset \tau'$ as $\tau''$ is the coarsest topology that makes all sim(., $Q$) continuous. □

It will be shown in this and in the next section that it is possible that $\tau \neq \tau''$, $\tau'' \neq \tau'$, and even $\tau \neq \tau'' \neq \tau'$ are possible, as well as $\tau = \tau'' = \tau'$!

Everett and Cater (1992) claimed the following results:

*Statement 1*

For any $Q \in QS$ and $t \in [0, 1]$, the set $U(Q, t) = \{D \in DS | \mathrm{sim}(D, Q) < t\} \in \tau$.

*Statement 2*

Let $\tau_1$ and $\tau_2$ be the retrieval topologies of two essentially equivalent retrieval systems. Assume that ($DS$, $\tau_2$) is compact, then $\tau_1 = \tau_2$.

Recall (Everett & Cater, 1992) that two retrieval systems ($DS$, $QS$, $\mathrm{sim}_1$) and ($DS$, $QS$, $\mathrm{sim}_2$) are said to be essentially equivalent if $\mathrm{sim}_1(D, Q) < \mathrm{sim}_1(E, Q)$ iff $\mathrm{sim}_2(D, Q) < \mathrm{sim}_2(E, Q)$. Essentially equivalent models retrieve all documents in the same order.

In Egghe and Rousseau (1997b), we have shown that these statements (Lemma 1 and Theorem 4 in Everett & Cater, 1992) are wrong. We briefly repeat the counterexamples constructed in Egghe and Rousseau (1997b) and add some new comments.

*A Counterexample to Statement 1.* Let $DS = \mathbf{N}$ (all natural numbers, including zero); $QS$ can be any nonempty set. Let $\mathrm{sim}_1$ be defined as follows: $\mathrm{sim}_1(D, Q) = \frac{1}{5}$ for all $Q \in QS$, except for one special query $Q_1$. For this special query $\mathrm{sim}_1(n, Q_1)$, $n \in \mathbf{N}$, is given by the following table:

| $n$ | $\text{sim}_1(n, Q_1)$ |
|---|---|
| 0 | 1/5 |
| 1 | 1/6 |
| 3 | 3/8 |
| 5 | 7/16 |
| 7 | 15/32 |
| and so on for the odd numbers | |
| 2 | 1/2 |
| 4 | 3/4 |
| 6 | 7/8 |
| and so on for the even numbers | |

Note that the similarity values for the odd numbers converge to $\frac{1}{2}$; the similarity values for the even numbers converge to 1.

Now the set $U(Q_1, \frac{1}{4}) = \{D \in DS | \text{sim}_1 (D, Q_1) < \frac{1}{4}\} = \{0\}$ does not belong to the retrieval topology $\tau_1$. Indeed, the sets of $\tau_1$ are the following:

- **N** and $\varnothing$ (the empty set);
- all natural numbers except zero;
- all natural numbers except zero less the $j$ smallest odd numbers ($j = 1, 2, \cdots$);
- all natural numbers except zero, less the odd numbers and the $j$ smallest even numbers ($j = 1, 2, 3, \cdots$).

Note that the set consisting of all even numbers $\{2, 4, 6, \cdots\}_{337}$ is NOT an open set for this retrieval topology.

This example also shows that the function $\text{sim}_1(., Q_1)$ is not continuous for the retrieval topology. Hence, this is also a case where the retrieval topology differs from the similarity topology.

*A Counterexample to Statement 2.* We consider the same retrieval system as before but—for the second similarity function—make a slight change to the first one. The similarity function $\text{sim}_2$ is everywhere equal to $\text{sim}_1$, except for the value in $(2, Q_1)$. We set $\text{sim}_2(2, Q_1)$ equal to $\frac{5}{8}$. It is now clear that the models ($DS$ = **N**, $QS$, $\text{sim}_2$) and ($DS$ = **N**, $QS$, $\text{sim}_2$) are essentially equivalent. Moreover, as $\text{sim}_i (0, Q)$ is $\frac{1}{5}$ for every $Q$, $i = 1, 2$, the set $DS$ = **N** is compact for the retrieval topology (the point zero plays the role of $D_0$ in the Proposition of Egghe and Rousseau (1997b). However, the two retrieval topologies do not coincide. For $\tau_2$ (the retrieval topology derived from $\text{sim}_2$), the set consisting of all even numbers $= \{D \in DS | \text{sim}_2(D, Q_1) > \frac{1}{2}\}$ is clearly open (an element of $\tau_2$). We noted before that this set is not open in $\tau_1$. Hence, the two topologies are not equal, which contradicts statement 2.

We consider now the following question: Can statements 1 and 2 be adapted so that they become true? The answer to this is yes, essentially by using $\tau''$ instead of $\tau$. This will be shown in Theorems 2.8.1 and 2.8.2. We first state a simple lemma and introduce a new notion.

*Lemma 2.6*

For any $Q \in QS$ and any $r \in$ **R,** the set

$$\{D \in DS | \text{sim}(D, Q) < r\}$$

belongs to the similarity topology $\tau''$.

*Proof.* This follows readily from the definition of $\tau''$ and the fact that the sets $]-\infty, r[$ are open in **R,** for every $r \in$ **R.** □

Lemma 2.6 shows that Lemma 1 in Everett and Cater (1992) is true for $\tau''$. Also, their Theorem 4 becomes true when using $\tau''$ instead of $\tau$; see theorem 2.8.2 further on.

*Definition 2.7*

We say that a retrieval system ($DS$, $QS$, sim) satisfies the maximum principle if $\forall r \geq 0$, $\forall Q \in QS$, the set

$$A = \{\text{sim}(D, Q) | D \in DS, \text{sim}(D, Q) \leq r\}$$

has a maximum, i.e., there exists $D_0 \in DS$ such that $\text{sim}(D_0, Q) = \max A$. An analogous definition can be given for the minimum principle.

*Examples.* We formulate some cases of retrieval systems satisfying Definition 2.7 (minimum as well as maximum principle).

(i) If the range of sim(., $Q$) is finite for every $Q \in QS$, then the system satisfies the minimum as well as the maximum principle. This is the case for the classical example where sim can only take values in $\{0, 1\}$. This is also the case for a finite document space.
(ii) If $DS$ is compact for a certain topology $S$ such that all sim(., $Q$) are continuous (e.g., $\tau''$), then the system also satisfies the minimum and the maximum principle. Indeed, under these conditions, sim($DS$, $Q$) is compact in **R** for every $Q \in QS$. Consequently, sim($DS$, $Q$) is closed and bounded, and therefore has a minimum and a maximum.

This brings us to Theorems 2.8.1 and 2.8.2.

*Theorem 2.8.1*

If ($DS$, $QS$, $\text{sim}_1$) and ($DS$, $QS$, $\text{sim}_2$) are essentially equivalent retrieval systems, and if one of the two satisfies the maximum principle, then their retrieval topologies are equal.

*Proof.* If one model satisfies the maximum principle then, by the fact that they are essentially equivalent, the other model does too (and the maxima are reached for the same document $D_0$). For every $Q \in QS$ and $r \in$ **R,** let $D_0 \in DS$ be such that $\text{sim}_1 (D_0, Q) = \max \{\text{sim}_1(E, Q);\ \text{sim}_1$

$(E, Q) \leq r\}$. Then the following expressions are equivalent:

$$D \in R_1(Q, r) = \{E \in DS | \mathrm{sim}_1(E, Q) > r\}$$

$$\Leftrightarrow \mathrm{sim}_1(D, Q) > \mathrm{sim}_1(D_0, Q)$$

$$\Leftrightarrow \mathrm{sim}_2(D, Q) > \mathrm{sim}_2(D_0, Q)$$

$$\Leftrightarrow D \in R_2(Q, \mathrm{sim}_2(D_0, Q))$$

The same argument can be given when the indices are interchanged. Since the sets $\{R_i(Q, r) | Q \in QS, r \in \mathbf{R}\}$, $i = 1, 2$ form a subbasis for the respective retrieval topologies, these topologies are the same. □

*Theorem 2.8.2*

If $(DS, QS, \mathrm{sim}_1)$ and $(DS, QS, \mathrm{sim}_2)$ are essentially equivalent retrieval systems, and if one of the two satisfies the minimum principle, and if one of the two satisfies the maximum principle, then their similarity topologies are equal.

*Proof.* Since the retrieval systems are essentially equivalent, they both satisfy the minimum and the maximum principle. In the same way as in the proof of Theorem 2.8.1, we can now show that, for every $Q \in QS$ and $r_1, r_2 \in \mathbf{R}$, $r_1 < r_2$:

$$\{E \in DS | r_1 < \mathrm{sim}_1(E, Q) < r_2\}$$

$$= \{E \in DS | \mathrm{sim}_2(D_1, Q) < \mathrm{sim}_2(E, Q) < \mathrm{sim}_2(D_0, Q)\}$$

where $D_0$ resp. $D_1$ are the documents featuring in the definition of the minimum resp. maximum principle of the first system. Hence

$$U_1(Q, r_1, r_2) = U_2(Q, \mathrm{sim}_2(D_1, Q), \mathrm{sim}_2(D_0, Q)).$$

Again, the same argument can be given with the indices interchanged and hence $\tau_1'' = \tau_2''$. □

The topology $\tau''$ is defined as the coarsest topology on $DS$ that makes all the functions $\mathrm{sim}(., Q) : DS \rightarrow \mathbf{R}$ continuous. Here, $\mathbf{R}$ has the usual Euclidean topology $\varepsilon$. Our next result shows that the retrieval topology can be defined in a similar way, but using a different topology on $\mathbf{R}$.

*Theorem 2.9*

Equip the real line $\mathbf{R}$ with the topology, $S$, generated by the open half-lines $]r, +\infty[$, $r \in \mathbf{R}$. Then the retrieval topology $\tau$ is the coarsest topology on $DS$ that makes all functions

$$\mathrm{sim}(., Q) : DS \rightarrow \mathbf{R}, S$$

continuous.

*Proof.* The functions $\mathrm{sim}(., Q)$ are continuous from $DS$ to $\mathbf{R}, S$ iff the sets $R(Q, r) = \mathrm{sim}(., Q)^{-1}(]r, +\infty[)$ are open in $DS$. □

One could also say that on $(DS, \tau)$ all sims are lower semi-continuous into $\mathbf{R}$, equipped with $\varepsilon$ (see Willard, 1970, p. 49, 7K). Also, the (pseudo-)metric topology $\tau'$ can be characterized through continuity properties of the similarity functions.

*Theorem 2.10*

The following assertions are equivalent:

(i) $\tau' = \tau''$;
(ii) The family $\{\mathrm{sim}(., Q) | Q \in QS\}$ is equicontinuous on $(DS, \tau'')$, where the functions range in $\mathbf{R}, \varepsilon$

For the proof, we refer to Appendix C.

*Corollary 2.11*

If $QS$ is finite, then $\tau' = \tau''$. Based on the above theorem and/or Theorem 2.5, we obtain the following easy results.

*Proposition 2.12*

The following assertions are equivalent:

(i) $\tau = \tau''$;
(ii) All functions $\mathrm{sim}(., Q)$ are continuous on $(DS, \tau)$, with range in $\mathbf{R}, \varepsilon$.

*Proposition 2.13*

The following assertions are equivalent:

(i) $\tau = \tau' = \tau''$;
(ii) The family $\{\mathrm{sim}(., Q); Q \in QS\}$ is equicontinuous on $(DS, \tau)$ with range in $\mathbf{R}, \varepsilon$.

## 3. Examples of Topologies on Document Spaces

### *3.1. The Counterexample to Statement 1*

We repeat that the counterexample to Statement 1 is an example for which $\tau = \tau''$. Since the similarity functions are clearly equicontinuous in $\tau''$, we conclude by Theorem 2.10 that $\tau'' = \tau'$.

### 3.2. The Vector Space Model of Everett and Cater (1992)

We briefly recall the structure of the vector space model as presented in Everett and Cater (1992) and point out some problems with this model. We will compare this vector model to another one in Section 3.3.

The Everett-Cater model starts by taking $I^n = [0, 1]^n$ for the document space, as well as for the query space, with the following similarity function:

$$\mathrm{sim}(D, Q) = \frac{\langle D, Q\rangle}{\|D\|_2\|Q\|_2} = \cos(D, Q) \qquad (13)$$

where $\langle .,.\rangle$ denotes the usual inner product, $\|.\|_2$ is the Euclidean norm and $\cos(D, Q)$ denotes the cosine of the angle between the lines $OD$ and $OQ$.

We see two problems with this model. Firstly, the model does not distinguish between documents on a straight line through the origin. Since documents in $[0, 1]^n$ are often obtained through a weighing process, it seems to us a waste of possibilities not to use these different weights. Secondly, the similarity functions are not defined in points $(D, Q)$ where at least one of the coordinates is zero. Moreover, it is impossible to extend sim in such a way that this function becomes continuous on $I^n$.

Everett and Cater solve these problems by taking $DS = QS = I^n\backslash\{0\}$ and introducing an equivalence relation $R$, where $DRE$ iff there is straight line through the origin containing both $D$ and $E$. Then $DS^* = (I^n\backslash\{0\})/R$ with the usual quotient norm (and hence quotient topology). For readers not familiar with quotient theory, we just note here that elements of $DS^*$ are straight half-lines through the origin (0 not included), and only the part inside $I^n$ is used. The similarity function 13 can be used unambiguously as above: If $D^* \in DS^*$ and $Q \in QS$, then

$$\mathrm{sim}(D^*, Q) = \frac{\langle D, Q\rangle}{\|D\|_2\|Q\|_2} \qquad (14)$$

for any representative $D \in D^*$. Also, for $D^*, E^* \in DS^*$, we can define:

$$\mathrm{sim}(D^*, E^*) = \frac{\langle D, E\rangle}{\|D\|_2\|E\|_2} \qquad (15)$$

for any representative $D \in D^*$, $E \in E^*$.

By taking quotients as above, we now obtain that the retrieval system separates points. Indeed, if $\mathrm{sim}(D^*, Q) = \mathrm{sim}(E^*, Q)$, $\forall Q \in QS$, then $\cos(D, Q) = \cos(E, Q)$ for any $D \in D^*$, $E \in E^*$, and hence $D^* = E^*$. This shows that here $\tau'$ is a metric topology. This leads to:

*Theorem 3.1 (Everett & Cater, 1992)*

$(DS^*, \tau') = (S, \varepsilon)$ with $S$ that part of the unit sphere that contains vectors with positive (or zero) coordinates and with $\varepsilon$ the Euclidean topology on $S$.

*Proof.* The homeomorphism is obtained via the function:

$$f : DS^* \to S : D^* \to D \cap S \quad \square \qquad (16)$$

*Theorem 3.2*

$\tau = \tau'' = \tau'$ on $DS^*$ and on $DS$.

*Proof.* The subbasic neighborhoods for an element $D \in DS$ or $D^* \in DS^*$, for $\tau$ as well as for $\tau'$, are of the form: An open cone with top 0 (not an element of $DS^*$) and $D^*$ as central ray. Hence $\tau = \tau'$, and thus, by Theorem 2.5, $\tau = \tau' = \tau''$. $\square$

*Corollary 3.3*

$DS^*$ with $\tau = \tau' = \tau''$, is compact.

*Proof.* This follows from Theorem 3.1 and Theorem 3.2.

*Corollary 3.4 (Everett & Cater, 1992)*

$DS$ with $\tau = \tau' = \tau''$ is compact.

*Proof.* The argument that proves that $(S, \varepsilon)$ is compact can be used on $DS$ if we use the function $f^{-1}$ (the inverse of function $f$ of Equation 16) and consider $D^*$ as a subset of $DS$. $\square$

*Note.* It is not true that $DS$ and $DS^*$ are homeomorphic! Indeed, $DS^*$ is a Hausdorff (or $T_2$) space, while $DS$ is not.

The vector space model that we will present in the next section is an alternative for the usual one. Its mathematical properties are less complicated and more interesting, as it will be defined on $I^n$ (not on a quotient space) and will give different topologies $\tau \neq \tau''$. In addition to this, it takes into account the different weights of vectors in $I^n$.

### 3.3. An Alternative for the Classical Vector Space Model

Using the cosine of the angle between vectors does not take into account the different weights given to each coordinate (representing, e.g., a keyword). In addition to this, documents can be situated close to each other, yet the similarity as measured by cosines can be relatively small too. Using the inner product between vectors takes care of these problems. In such a model, a query gives weights to keywords. These weights can be interpreted as minimum requirements: Documents with higher weights (for that particular keyword) will just score better.

We will now formalize this. Let $DS = QS = I^n = [0, 1]^n$, $n \in \mathbf{N}_0$. For every $D \in DS$, $Q \in QS$:

$$\mathrm{sim}(D, Q) = \langle D, Q\rangle = \|D\|_2\|Q\|_2 \cos(D, Q). \qquad (17)$$

The following results describe the topologies of this retrieval model.

*Theorem 3.5*

(i) The function $\langle . , . \rangle : \times QS \rightarrow \mathbf{R}_+; (D, Q) \rightarrow \langle D, Q \rangle$ is continuous for the norm topologies on $DS = QS = I^n$.
(ii) The retrieval system ($DS$, $QS = DS$, sim), with sim defined in Equation 17 separates the points of $DS$.

The proof of these results is well-known and is omitted.

*Theorem 3.6*

In this model, $\tau' = \varepsilon$, the Euclidean topology on $I^n$. Hence, here $DS$ with any of the topologies $\tau$, $\tau'$, or $\tau''$ is compact.

*Theorem 3.7*

For this model, $\tau'' = \tau'$, hence $DS$ with this topology is a $T_2$ space.

*Theorem 3.8*

For this model, $(DS, \tau)$ is a $T_0$-space, that is, not a $T_1$-space. Consequently $\tau \neq \tau''$.

The proof of these theorems can be found in Appendix D.

### 3.4. Another Simple Example of a Document System for Which $\tau \neq \tau'' = \tau'$

Let $DS = \{a, b, c, d\}$, $QS = \{e\}$ and define

$$\text{sim}(a, e) = 0.5 = \text{sim}(d, e)$$

$$sim(b, e) = 1, \text{sim}(c, e) = 0$$

Then:

$$\tau = \{\emptyset, DS, \{a, b, d\{, \{b\}\}$$

while

$$\tau' = \tau'' = \{\emptyset, DS, \{b\}, \{c\}, \{a, d\}, \{b, c\}, \{a, b, d\}, \{a, c, d\}\} \neq \mathcal{P}(DS),$$

the discrete topology. Note that the equality between $\tau'$ and $\tau''$ illustrates Corollary 2.11, as here $QS$ is a finite set. This retrieval system does not separate points.

### 3.5. Three Examples for Which $\tau = \tau' = \tau'' = \mathcal{P}$ (DS)

Example 3.2 gave an example for which $\tau = \tau' = \tau''$, but different from the discrete topology, since they were homeomorphic with the Ecuclidean topology on $S$. Three examples will follow where all topologies are equal to $\mathcal{P}$ ($DS$), the discrete topology on $DS$.

*3.5.1.* $DS = QS$ any set

$$\text{sim}(D, Q)\begin{cases} = 0, & D \neq Q \\ = 1, & D = Q \end{cases} \tag{18}$$

Hence, in this IR-model only exact matches are allowed! So, $\forall Q \in QS, \forall r \in \mathbf{R}, R(Q, r) = \emptyset$ if $r \geq 1$ and $= \{Q\}$ if $r < 1$. Hence, $\tau = \mathcal{P}$ ($DS$). By Theorem 2.5, $\tau \subset \tau'' \subset \tau'$. Hence, $\tau = \tau'' = \tau' = \mathcal{P}$ ($DS$).

Note that the metric of $\tau'$ is the discrete metric

$$d(D, E)\begin{cases} = 1 & \text{if } D = E \\ = 0 & \text{if } D \neq E \end{cases} \tag{19}$$

Discrete metrics yield discrete topologies but not conversely, as the next two examples show.

*3.5.2.* The next example deals with documents that are ordered according to "similarity." This is an index approach and cannot always be used in practice. However, if topics can be ordered this way (e.g., for aspects on which a linear order applies, such as distance, temperature, . . .) the example is useful. Let

$$DS = \{D_1, . . ., D_n\} = QS, n \in \mathbf{N} \text{ fixed.}$$

Define

$$\text{sim}(D_i, D_j) = \frac{i + j}{2 \max(i, j)}. \tag{20}$$

Intuitively, this means that if $i$ and $j$ are far apart, then $\text{sim}(D_i, D_j)$ is small, and if $i \approx j$, then $\text{sim}(D_i, D_j) \approx 1$. In any case, $\text{sim}(D_i, D_j) = 1$, for every $i \in \{1, . . . , n\}$. We have the following results:

*Theorem 3.9*

$\forall j = 1, . . . , n, \ \forall r > 0$:

$$R(D_j, r) = \{D_i \in DS | i \in \{1, . . . , n\} \cap \left]2jr - j, \frac{j}{2r - 1}\right[ \tag{21}$$

which forms a subbasis for $\tau$.

*Theorem 3.10*

Let $d$ be the metric of $\tau'$. Then, $\forall i, k = 1, \ldots, n$:

$$d(D_i, D_k) = \frac{|i-k|}{2\max(i, k)}. \tag{22}$$

*Corollary 3.11* $\tau = \tau'' = \tau', = \mathcal{P}(DS)$.

*For the proofs, we refer the reader to Appendix E.*

*A similar example, but with a completely different similarity function, now follows.*

3.5.3. $DS = QS = \{D_1, \ldots, D_n\}$ $n \in \mathbf{N}$ fixed. Define

$$\mathrm{sim}(D_i, D_j) = \frac{2}{\pi}\,\mathrm{Arctan}\left(\frac{1}{|i-j|}\right). \tag{23}$$

Note again that $\mathrm{sim}(D_i, D_i) = 1$, $\forall i = 1, \ldots, n$.

*Theorem 3.12*

$\forall j = 1, \ldots, n$, $\forall r > 0$.

$$R(D_j, r) = \left\{D_i \in DS \mid i \in \{1, \ldots, n\} \cap \left]j - \frac{1}{\tan\frac{\pi r}{2}}, j + \frac{1}{\tan\frac{\pi r}{2}}\right[\right\} \tag{24}$$

which forms a subbasis for $\tau$.

*Theorem 3.13*

Let $d$ be the metric of $\tau'$. Then, $\forall i, k = 1, \ldots, n$:

$$d(D_i, D_k) = \frac{2}{\pi}\,\mathrm{Arctan}|i-k|. \tag{25}$$

*Corollary 3.14*

$$\tau = \tau'' = \tau' = \mathcal{P}(DS).$$

For the proofs, we refer the reader to Appendix F.

### *3.6. An Example for Which $\tau \neq \tau' \neq \tau''$*

This final example shows that $\tau \neq \tau' \neq \tau''$ is possible. In view of Theorem 2.10 and its Corollary 2.11, $QS$ must be an infinite space. This means that we must admit an infinite number of possible queries.

Let $DS = QS = [0, +\infty[$, with $\mathrm{sim}(D, Q) = D \cdot Q$ (the simple product of the numbers $D$ and $Q$). Let $(D_n)_{n=1,\ldots,\infty}$ be a strictly increasing sequence in $DS$, converging to $D$ in the usual, Euclidean norm on $[0, +\infty[$ (i.e., the absolute value). Consequently, $\lim_{n\to\infty} D_n = D$ in $\tau''$ since, for every $Q \in QS$:

$$\mathrm{sim}(D_n, Q) = D_n \cdot Q \to D \cdot Q = \mathrm{sim}(D, Q).$$

Now, $D_n$ does not tend to $D$ in $\tau'$ since

$$d'(D_n, D) = \sup_{Q \in QS}(\min(1, |\mathrm{sim}(D_n, Q) - \mathrm{sim}(D, Q)|))$$

which is equal to 1 if $D_n \neq D$ and is equal to 0 if $D_n = D$. As here $D_n$ is never equal to $D$, $d'(D_n D)$ is always equal to 1.

From this, it follows that $\tau' = \mathcal{P}(DS) \neq \tau''$. We still have to prove that $\tau \neq \tau''$. We know already that $\tau''$ yields a $T_2$-space, since $\tau''$ coincides here with the Euclidean topology on $[0, +\infty[$. (This follows from the fact that a sequence $D_n \to D$ in $|\cdot|$ iff $D_n Q \to D \cdot Q$ for every $Q \in [0, +\infty[$.) Now, $(DS, \tau)$ is not even a $T_1$-space: For every $D \in DS$ and $E = \alpha D$, $\alpha > 1$, and for every $\tau$-neighborhood $U$ of $D$, $\exists n \in \mathbf{N}$ and $R(F_j, r_j)$ such that

$$D \in \cap_{j=1}^{n} R(F_j, r_j) \subset U.$$

Hence, $\mathrm{sim}(D, F_j) = D \cdot F_j > r_j$, for every $j = 1, \ldots, n$. Since $D \cdot F_j \in [0, +\infty[$ also $\mathrm{sim}(E, F_j) = \alpha D \cdot F_j > r_j$, which implies that $E \in U$. This shows that $(DS, \tau)$ is not a $T_1$-space.

In fact, if $[0, +\infty[$ is equipped with the topology $D$ inherited from $S$ on $\mathbf{R}$ (cf. Theorem 2.9), then $\tau = D$ (restricted to $[0, +\infty[$). We conclude:

$\tau$ is the topology inherited from $S$ on $\mathbf{R}$
$\tau''$ is the Euclidean topology
$\tau'$ is the discrete topology on $DS$,

and all three of them are different.

## 4. Boolean Information Retrieval as a Subsystem of Any Topological IR-System

In Cater (1986) and Everett and Cater (1992), the Boolean IR-model is shown to be an example of topological retrieval, in the sense that specially defined similarity values yield traditional Boolean retrieval. In this section, we will show that Boolean retrieval can be introduced in any IR-model ($DS$, $QS$, sim) without having to specify the value of a similarity function for Boolean combinations of elementary queries. The Boolean model we will present has the further advantage that it presents a major clarification of why the introduction of topologies in theoretical IR-models is essential.

As defined in Section 1, let, for every $Q \in QS$,

$$\mathrm{ret}_\tau(Q) = \{R(Q, r) | r \in \mathbf{R}\}$$

and

$$\mathrm{ret}_{\tau''}(Q) = \{U(Q, r_1, r_2) | r_1, r_2 \in \mathbf{R}, r_1 < r_2\}$$

be the retrieval sets of $Q$ w.r.t. the topology $\tau$ (resp. $\tau''$). Later in this section, we will focus on $\mathrm{ret}_\tau(Q)$, denoting this set simply as $\mathrm{ret}(Q)$. (Exactly the same reasoning can be made for $\tau''$).

The set $QS$ consists of elementary queries, which means that usually logical combinations of queries do not belong to $QS$, although they are not excluded either. $\mathrm{ret}(Q)$ consists of all possible results when $Q$ is used as a query. The only difference between the use of the topologies $\tau$ and $\tau''$ is that we require "minimum" (threshold) values for sim($\tau$ case), or "close" values ($\tau''$ case).

Recall also that the sets in all $\mathrm{ret}_\tau(Q)$, $Q \in QS$, form a subbasis for $\tau$. This means that

$$B = \left\{ \bigcap_{i=1}^{n} R(Q_i, r_i) \,\middle|\, Q_i \in QS, r_i \in \mathbb{R}, n \in \mathbb{N}_0 \right\}$$

is a basis for $\tau$ (similarly, a basis for $\tau''$ can be built using the sets $U(Q, r_1, r_2)$). An open set A in $\tau$ is then any union (finite or infinite) of sets in $B$:

$$A = \bigcup_{j \in J} \left( \bigcap_{i=1}^{n_j} R(Q_{ij}, r_{ij}) \right)$$

where $J$ is a finite or infinite index set. From now on, we will assume that $DS$ and $QS$ are finite. Note that then $\tau'' = \tau'$ (Theorem 2.10 and Corollary 2.11). So, we have only $\tau$ and $\tau''$ to consider, and, as stated before, we will focus on $\tau$. In this case, a general open set A in $\tau$ has the form

$$A = \bigcup_{j=1}^{m} \left( \bigcap_{i=1}^{n_j} R(Q_{ij}, r_{ij}) \right). \tag{26}$$

*Definition 4.1*

Let $Q_1, \ldots, Q_n$ be $n$ elements of $QS$. We introduce, symbolically, an element $\otimes_{i=1}^{n} Q_i$ and define

$$\mathrm{ret}\left( \bigotimes_{i=1}^{n} Q_i \right) = \left\{ \bigcap_{i=1}^{n} R(Q_i, r_i) \,\middle|\, r_i \in \mathbf{R} \right\} \tag{27}$$

We call $\otimes_{i=1}^{n} Q_i$ the Boolean AND applied to the elementary queries $Q_1, \ldots, Q_n \in QS$. By definition, this Boolean AND retrieves sets in $\mathrm{ret}(\otimes_{i=1}^{n} Q_i)$. We repeat that $\otimes_{i=1}^{n} Q_i$ is not necessarily an element of $QS$. Similarly, we define a Boolean OR.

*Definition 4.2*

Let $Q_1, \ldots, Q_n$ be $n$ elements of $QS$. Consider, symbolically, an element $\oplus_{i=1}^{n} Q_i$ and define

$$\mathrm{ret}\left( \bigoplus_{j=1}^{m} Q_j \right) = \left\{ \bigcap_{j=1}^{m} R(Q_j, r_j) | r_j \in \mathbb{R} \right\}. \tag{28}$$

The set $\oplus_{j=1}^{m} Q_j$ is called the Boolean OR applied to the elementary queries $Q_1, \ldots, Q_n \in QS$. By definition, this Boolean OR retrieves sets in $\mathrm{ret}(\oplus_{j=1}^{m} Q_j)$.

Of course, we can as well define Boolean ANDs and Boolean ORs with respect to the topology $\tau''$. We suppose that it is clear in every IR-search whether we want to work with $\tau$ (thresholds) or $\tau''$ (close matches).

Based on Definitions 4.1 and 4.2, we can easily define more arbitrary Boolean queries, based on elementary queries in $QS$.

*Definition 4.3*

*Let*$(Q_{ij})_{i=1}^{n_j}{}_{j=1}^{m}$ be an array of queries in $QS$. The Boolean query (not necessarily in $QS$) $\otimes_{j=1}^{m} (\otimes_{i=1}^{n_j} Q_{ij})$ is defined through its retrieval:

$$\mathrm{ret}\left( \bigotimes_{j=1}^{m} \left( \bigotimes_{i=1}^{n_j} Q_{ij} \right) \right)$$

$$= \left\{ \bigcup_{j=1}^{m} \left( \bigcap_{i=1}^{n_j} R(Q_{ij}, r_{ij}) \right) \,\middle|\, r_{ij} \in \mathbb{R} \right\} \tag{29}$$

and similarly when $\oplus$ and $\otimes$ (hence $\cup$ and $\cap$) are interchanged. We now need the following lemma:

*Lemma 4.4 (Dugundji, 1966, p. 25)*

Let $\{B_\alpha | \alpha \in A\}$ be a family of sets and assume that $\{A_\lambda; \lambda \in \Lambda\}$ is a partition of A (hence each $A_\lambda$; $\lambda \varepsilon \neq \emptyset$). Let $T = \Pi_{\lambda \in \Lambda} A_\lambda$.
Then

$$\bigcap_{\lambda \in \Lambda} \left( \bigcup_{\alpha \in A_\lambda} B_\alpha \right) = \bigcup_{t \in T} \left( \bigcap_{\lambda \in \Lambda} B_{t(\lambda)} \right) \tag{30}$$

where $t(\lambda) \in A_\lambda$ and $t = (t(\lambda))_{\lambda = \Lambda}$.

By taking complements, one also has:

$$\bigcup_{\lambda \in \Lambda} \left( \bigcap_{\alpha \in A_\lambda} B_\alpha \right) = \bigcap_{t \in T} \left( \bigcup_{\lambda \in \Lambda} B_{t(\lambda)} \right). \tag{31}$$

In plain terms: Any introduction of unions of sets $A_{ij}$ can be interpreted as a union of intersections of the same sets (but in another order) and vice versa.

We have now reached the following important result, yielding a major reason why topologies on IR-systems are useful.

*Theorem 4.5*

For any IR-model ($DS$, $QS$, sim) with finite $DS$ and $QS$, we have the following equalities:

a) The topology $\tau$ is equal to the set of all possible Boolean retrievals, based on elementary queries $Q$ in $QS$, using thresholds;
b) the topology $\tau''$ is equal to the set of all possible Boolean retrievals, based on elementary queries $Q$ in $QS$, using close matches.

*Proof.* The principal elements of the proof have already been outlined. An arbitrary Boolean query can be denoted as a combination of AND and OR Boolean queries in any order, which can be denoted as in Equation 29 or with $\cup$ and $\cap$ interchanged (cf. Definition 4.3). By lemma 4.4, its set of retrievals can be written in the form:

$$\left\{ \bigcup_{j=1}^{m} \left( \bigcap_{i=1}^{n_u} \mathbf{R}(Q_{ij}, r_{ij}) \right) \middle| r_{ij} \in \mathbf{R} \right\}.$$

Letting $m$ and $n_j$ vary over all possible natural numbers and $Q_{ij} \in QS$, we see, by Equation 26, that the set of all Boolean retrievals using thresholds, is nothing but the topology $\tau$. Similarly, the set of all Boolean retrievals using close matches is nothing but the topology $\tau''$. □

The following result gives a relation between the usual OR relation and our somewhat more general use of the Boolean OR.

*Proposition 4.6*

If $Q_1$ and $Q_2$ are single attribute queries, if also $Q = Q_1$ OR $Q_2$ belongs to $QS$, and if sim($D$, $Q$) is defined as max(sim($D$, $Q_1$), sim($D$, $Q_2$))—as in the classical Boolean or in the fuzzy set case-then (using thresholds)

$$\text{ret}(Q) \subset \text{ret}\left( \bigoplus_{i=1}^{2} Q_i \right).$$

*Proof.* Consider the set $\{D \in DS | \text{sim}(D, Q) > r\}$, an element of ret($Q$). As sim($D$, $Q$) = max(sim($D$, $Q_1$), sim($D$, $Q_2$)), this set is equal to $\{D \in DS | \max(\text{sim}(D, Q_1), \text{sim}(D, Q_2)) > r\}$.
$= \{D \in DS | \text{sim}(D, Q_1) > r, \text{ or}, \text{sim}(D, Q_2) > r\}$
$= \{D \in DS | \text{sim}(D, Q_1) > r\} \cup \{D \in DS | \text{sim}(D, Q_2) > r\}$
$= R(Q_1, r) \cup R(Q_2, r) \in \text{ret}(Q_1 \otimes Q_2)$.

A similar result can be shown for any finite disjunctive form, and for any finite conjunctive form. The Boolean NOT will be studied in a future article.

## 5. Summary

In this article, Everett and Cater's retrieval and pseudometric topologies are examined. Counterexamples to two statements are given, and results correcting the original statements are presented. These corrections lead to the introduction of a new topology, namely the similarity topology. This new topology satisfies the interesting property of making the similarity functions continuous (stable).

Several examples are presented, among them a modified vector space model. Finally, the article shows that the retrieval and the similarity topology can be considered as the sets of retrievals of arbitrary Boolean AND/OR queries.

## Appendix A: Generalities on Topological Spaces

Let $X$ denote a set. Denote by $\mathcal{P}(X)$ the set of all subsets of $X$. A *topology* $\tau$ on $X$ is a subset of $\mathcal{P}(X)$ such that:

(i) Any union of elements in $\tau$ belongs to $\tau$;
(ii) any finite intersection of elements in $\tau$ belongs to $\tau$;
(iii) $\varnothing$ (the empty set) and $X$ belong to $\tau$.

The element of $\tau$ are called *open* sets. The couple ($X$, $\tau$) is called a *topological space.*

Given two topologies $\tau_1$ and $\tau_2$ on $X$, we say that $\tau_1$ is *weaker* (or *smaller,* or *coarser*) than $\tau_2$, if $\tau_1 \subset \tau_2$. We then also say that $\tau_2$ is *stronger* (or *larger,* or *finer*) than $\tau_1$.

A set $F \subset X$ is called *closed* if its complement $F^c \in \tau$. If $A \subset X$, the *closure* of A in $X$, w.r.t. $\tau$, denoted by $\bar{A}$, is the set

$$\bar{A} = \cap \{B \subset X | B \text{ is closed and } A \subset B\}.$$

If it is not clear that the closure is w.r.t. $\tau$, we denote $\bar{A}^\tau$.

If ($X$, $\tau$) is a topological space, a *base* for $\tau$ is a collection $\mathcal{B} \subset \tau$, such that

$$\tau = \left( \bigcup_{B \in \mathcal{C}} B \middle| \mathcal{C} \subset \mathcal{B} \right\}.$$

A *subbase* for $\tau$ is a collection $\mathcal{C} \subset \tau$, such that the collection of all finite intersections of elements of $\mathcal{C}$ forms a base for $\tau$. Any collection of subsets of a set $X$ is a subbase for some topology on $X$. This follows from the definition of $\tau$. This assertion is not true for a base!

Note that $\mathcal{B}$ is a basis for $\tau$ iff, whenever $G \in \tau$ and $X \in G$, there is a $B \in \mathcal{B}$ such that $x \in B \subset G$.

Let $x \in X$. We say that $U \subset X$ is a *neighborhood* of $x$ if there exists $G \in \tau$, such that $x \in G \subset U$. We denote by $V_\tau(x)$ the collection of all neighborhoods of $x \in X$ and it is called a *neighborhood system.* A *neighborhood base* at $x \in X$ is a collection $B_\tau(x) \subset V_\tau(x)$ such that each $U \in V_\tau(x)$ contains an element $V \in B_\tau(x)$, i.e., $V_\tau(x)$ *is determined by* $B_\tau$ as

$$V_\tau(x) = \{U \subset X | V \subset U, \ \exists V \in B_\tau(X)\}$$

The elements of $B_\tau(x)$, once chosen, are called *basic neighborhoods.*

Let $(X, \tau)$ be a topological space. We say that it is a $T_0$ *space,* if for every $x, y \in X$, $x \neq y$, there exists a neighborhood of one of these points not containing the other. We say that it is a $T_1$ *space,* if for every $x, y \in X$, $x \neq y$, there exist neighborhoods of each of these points not containing the other point.

Let $(X, \tau)$ be a topological space. We say that $(X, \tau)$ is a *Hausdorff* (or $T_2$) *space,* if for every $x, y \in X$, $x \neq y$, there exists $G, H \in \tau$ such that $x \in G$, $y \in H$, and $G \cap H = \varnothing$. Equivalently, if there exists $U \in V_\tau(x)$, $V \in V_\tau(y)$, such that $U \cap V = \varnothing$.

Clearly, $T_2 \Rightarrow T_1 \Rightarrow T_0$. In this article, we will have examples showing that the converse is not true.

Let $X$ be a set and

$$d : X \times X \rightarrow \mathbf{R}_+$$

$$(x, y) \rightarrow d(x, y)$$

a function such that

(i) $d(x, x) = 0$
(ii) $d(x, y) = d(y, x)$
(iii) $d(x, y) \leq d(x, z) + d(z, y)$ (triangle inequality)

for every $x, y, z \in X$. Then $d$ is called a *pseudo-metric* on $X$ and $(X, d)$ is called a *pseudo-metric space.* Every pseudo-metric space can be considered as a topological space as follows:

(a) Define the "open balls," $\forall\ x \in X$, $\forall \varepsilon > 0$ $B(x, \epsilon) = \{y \in X | d(x, y) < \epsilon\}$.
(b) Define sets $A \subset X$ open iff for every $x \in A$, there is a $B(x, \varepsilon) \subset A$. It can be shown (cf. Willard, 1970, 2.6, p. 18) that the open sets defined above form a topology on $X$, the so-called *pseudo-metric topology* on $X$.

If (i) above can be reformulated as
$(i)'$ $d(x, y) = 0$ iff $x = y$
then we call $d$ a *metric,* $(X, d)$ a *metric space,* and the topology that it generates, the *metric topology.*

If $X$ is a vector space over $\mathbf{R}$, we can even go futher. We say that

$$\|\bullet\| : X \rightarrow \mathbf{R}_+$$

is a *pseudo-norm* on $X$ if

(i) $\|\vec{0}\| = 0$ ($\vec{0}$ is the zero vector in $X$)
(ii) $\|\alpha x\| = |\alpha|\ \|x\|$, $\forall \alpha \in \mathbf{R}$
(iii) $\|x + y\| \leq \|x\| + \|y\|$

for all $x, y \in X$. $\|\cdot\|$ is called a *norm* if (i) can be replaced by

$$(i)'\ \|x\| = 0 \text{ iff } x = \vec{0}$$

for all $x \in X$.

$X$, $\|\cdot\|$, accordingly, is called a (pseudo-)normed space. If we define

$$d(x, y) = \|x - y\|$$

then it can be shown that $d$ is a (pseudo-)metric.

*Example*

$\mathbf{R}^n$, $\|\bullet\|$, with, if $x = (x_1,, x_n) \in \mathbf{R}_n$

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p},$$

for $p \geq 1$.

$\|\cdot\|_p$ is called the *Minkowski norm* and, if $p = 2$, $\|\cdot\|_2$ is called the *Euclidean norm,* **d** derived from $\|\cdot\|_2$ is called the *Euclidean metric,* and its topology the *Euclidean topology,* denoted by $\varepsilon$. For $p = \infty$, we define

$$\|x\|_\infty = \max_{i=1,\dots,n} |x_i|.$$

Also, $\|\cdot\|_\infty$ is a norm on $\mathbf{R}^n$. All norms $\|\cdot\|_p$ $(1 \leq p \leq \infty)$ are *equivalent,* by which we mean that they all induce the same topology (in this case, $\varepsilon$). The same definition goes for equivalent pseudo-metrics.

$\|\cdot\|_2$ has one special feature, however. Define, for $x, y \in \mathbf{R}^n$

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i$$

the so-called *inproduct* in $\mathbf{R}^n$. This inproduct is used in this article. It makes $\mathbf{R}^n$ into an *inproduct space.* Every inproduct space $X$ is a normed space via the definition

$$\|x\| = \sqrt{\langle x, x\rangle}$$

for every $x \in X$.

The inproduct satisfies the so-called inequality of Cauchy-Schwarz: $\forall x, y, \in X$

$$\|\langle x, y\rangle\| \geq \|x\| \bullet \|y\|.$$

Let $(X, \tau)$ be a topological space and $Y \subset X$. Then

$$\tau_y = \{G \cap Y | G \in \tau\}$$

is a topology on $Y$, the *subspace* topology induced on $Y$ by $\tau$.

*Example*

$I^n \subset \mathbf{R}^n$ where $I = [0, 1]$.

We can take the restriction of $\|\cdot\|_2$ (or any other norm on $\mathbf{R}^n$) to $I^n$. This yields the Euclidean topology on $I^n$, being the subspace topology of $I^n$, inherited from the Euclidean topology on $\mathbf{R}^n$.

Let $(X, \tau)$ and $(Y, \tau')$ be two topological spaces and let $f: X \to Y$ be a function. We say that $f$ is *continuous,* if for every $B \in \tau', f^{-1}(B) \in \tau$. We say that $f$ is *open* if $G \in \tau$ implies $f(G) \in \tau'$. If $f: X \to Y$ is a bijection that is open and continuous, then we say that $(X, \tau)$ *and* $(Y, \tau')$ are *homeomorphic,* and we denote $(X, \tau) \cong (Y, \tau')$. Topologically spoken these spaces are indistinguishable.

Let $(X, \tau)$ be a topological space. We say that $\tau$ is *compact* if every open cover of $X$ has a finite subcover.

*Eample*

$[0, 1]$ is compact by $]0, 1[$ is not $(]1/n, 1[, n \in \mathbf{N}$ is an open cover of $]0, 1[$ without a finite subcover).

A set in $\mathbf{R}^n$ (with the Euclidean topology $\varepsilon$) is compact iff it is closed and bounded.

Let $f_j : (X, \tau) \to \mathbf{R}$ be a family of functions ($j \varepsilon J$, where the index set $J$ can be any, denumerable or non-denumerable, set). We say that the family $(f_j)_{j \in J}$ forms an equicontinuous family in $x \in X$, if for every $\varepsilon > 0$, there is a neighborhood $U$ of $x$, independent of $j \in J$, such that $y \in U$ implies that for every $j \in J : |f_j(x) - f_j(y)| < \varepsilon$. *The family* $f_j)_{j \in j}$ is said to equicontinuous on $(X, \tau)$ iff it is equicontinuous in every point $x$ of $X$. Note that every finite family of continuous functions is an equicontinuous family. Moreover, if the topology $\tau = \mathcal{P}(X)$, the discrete topology, then any family is equicontinuous.

## Appendix B: Proof of Theorem 2.1

*Theorem 2.1*

Let $d$, $d'$, and $d''$ be as defined in Equations 9, 11, and 12. Then

(i) the (pseudo-)metrics $d'$ and $d''$ are equivalent;
(ii) if $d$ exists, is is equivalent with $d'$ and $d''$.

*Proof.* For every $\varepsilon > 0$ and $D \in DS$, we denote by $B(D, \varepsilon)$, $B'(D, \varepsilon)$, *and* $B''(D, \varepsilon)$, the $d$- (resp. $d'$ and $d''$) $\varepsilon$-open ball around $D$.

(i) It is clear that for every $D, E \in DS$, $d'(D, E) \leq d''(D, E)$, hence, $B''(D, \varepsilon) \subset B'(D, \varepsilon)$ for every $\varepsilon > 0$ and $D \in DS$. Let now $\varepsilon > 0$ and $D \in DS$ be given. Set $\varepsilon' = \varepsilon/(2+\varepsilon)$. Then $d'(D, E) < \varepsilon'$ implies

$$\frac{\rho_Q(D, E)}{1 + \rho_Q(D, E)} < \varepsilon'$$

for every $Q \in QS$. Hence, $6_Q(D, E) < \varepsilon/2$ for every $Q \in QS$. So, $d''(D, E) < \varepsilon$ and, hence $B'(D, \varepsilon') \subset B''(D, \varepsilon)$. This shows that $d'$ and $d''$ generate the same topology on $DS$, denoted as $\tau'$.

(iii) Assume now that $d$ exists. Since for every $D, E \in DS : d'(D, E) \leq d(D, E)$, we see that $B(D, \varepsilon) \subset B'(D, \varepsilon)$ for every $\varepsilon > 0$ and $D \in DS$. Let now $B(D, \varepsilon)$ be an arbitrary open $\varepsilon$-ball around $D$. Let $\varepsilon' = \varepsilon/(2 + \varepsilon) > 0$, then we have for every $E \in B'(D, \varepsilon')$ that

$$d'(D, E) = \sup_{Q \in QS} \frac{\rho_Q(D, E)}{1 + \rho_Q(D, E)} < \varepsilon'.$$

This implies that, for every $Q \in QS$, $\varrho_Q(D, E) < \varepsilon/2$. Hence, $d(D, E) < \varepsilon$, which shows that $B'(D, \varepsilon) \subset B(D, \varepsilon)$. This proves part (ii) of the theorem. □

## Appendix C: Proof of Theorem 2.10

*Theorem 2.10*

The following assertions are equivalent:

(i) $\tau' = \tau''$
(ii) The family $\{sim(., Q) | Q \in QS\}$ is equicontinuous on $(DS, \tau'')$, where the functions range in $R$, $\mathscr{E}$.

*Proof:* $(ii) \Rightarrow (i)$. Let $(D_i)_{i \in I}$ be a net in $DS$, convergent to $D \in DS$ for $\tau''$. Since the set $\{\text{sim}(.,Q) | Q \in QS\}$ is equicontinuous on $\tau''$ into $\mathbf{R}, \mathscr{E}$, we have that $\forall \varepsilon > 0$, $\exists i_0(\varepsilon) \in$ I (independent of $Q \in QS$), such that $\forall i \geq i_0(\varepsilon)$

$$|\text{sim}(D_i, Q) - \text{sim}(D, Q)| < \frac{\varepsilon}{2}$$

$\forall Q \in QS$. Hence, $\forall Q \in QS$

$$\frac{|\text{sim}(D_i, Q) - \text{sim}(D, Q)|}{1 + |\text{sim}(D_i, Q) - \text{sim}(D, Q)|} < \frac{\varepsilon}{2}.$$

Hence,

$$\sup_{Q \in QS} \frac{|\text{sim}(D_i, Q) - \text{sim}(D, Q)|}{1 + |\text{sim}(D_i, Q) - \text{sim}(D, Q)|} < \varepsilon$$

so $d'(D_i, D) < \varepsilon$. Consequently, $(D_i)_{i \in I}$ converges to $D$ in $\tau'$ (by using the definition of $\tau'$). Hence, $\tau' = \tau''$.

*Proof. (i) ⇒ (ii).* Let the net $(D_i)_{i \in I}$ be convergent to $D$ on $DS$ equipped with the topology $\tau'' = \tau'$. Hence, $\forall \varepsilon > 0$, $\forall i_0)\varepsilon) \in I$, such that $\forall i \geq i_0(\varepsilon)$, $i \in I$:

$$d'(D_i, D) < \frac{\varepsilon}{1 + \varepsilon}.$$

From this, it follows that, $\forall Q \in QS$

$$| \operatorname{sim}(D_i, Q) - \operatorname{sim}(D, Q)| < \varepsilon$$

(hence, $\forall i \geq i_0(\varepsilon)$, independent of $Q$). Since this goes for every net $(D_i)_{i\in I}$ that converges in $\tau''$, we have that the set

$$\{\operatorname{sim}(., Q) | Q \in QS\}$$

is equicontinuous on $(DS, \tau'')$. □

## Appendix D: Proof of Theorems 3.6, 3.7, and 3.8

*Theorem 3.6*

$\tau' = \varepsilon$. Hence $\tau$, $\tau''$, $\tau'$ are compact.

*Proof.* Let $|\|\cdot\||$ be defined by, for $D \in DS$

$$|\|D\|| = \sup_{\|Q\|_\infty \leq 1} |\langle Q, D\rangle|$$

$$= \sup_{Q\in QS} |\langle Q, D\langle|$$

(since $D \in I^n$, we can indeed restrict ourselves to vectors $Q$ with positive coordinates). Hence, $d$ defined by

$$d(D, E) = |\|D - E\||$$

yields the metric topology $\tau'$. Furthermore, $d'$ defined by

$$d'(D, E) = \|D - E\|_2$$

with

$$\|D\|_2 = \sup_{\|Q\|_2\leq 1} |\langle Q, D\rangle|$$

yields the Euclidean metric. Since $\|\cdot\|_\infty$ and $\|\cdot\|_2$ are equivalent (see Appendix A), and since $\langle \alpha Q, D\rangle = \alpha\langle Q, D\rangle$ for every $\alpha \in \mathbf{R}$, it follows that $d$ and $d'$ are equivalent and hence, $\tau' = \varepsilon$.

Since $\tau \subset \tau'' \subset \tau' = \varepsilon$, and since $\varepsilon$ is compact, $\tau$, $\tau''$, and $\tau'$ also are. □

*Theorem 3.7*

$\tau' = \tau''$, hence they are $T_2$-spaces.

*Proof.* In view of Theorem 2.10, it suffices to prove that the set

$$\{ \langle .,Q\rangle | Q \in QS\}$$

is Euclidean on $(I'', \tau'')$. For this, it suffices to prove that

$$\sup_{Q\in QS} \|\langle ., Q\rangle\| < \infty$$

(see, e.g., Wilansky, 1970, p. 291, ex. 201).

Now

$$\sup_{Q\in QS} \|\langle ., Q\rangle\| = \sup_{Q\in QS} \sup_{\|D\|_2\leq 1} |\langle D, Q\rangle|$$

$$\leq \sup_{Q\in QS} \sup_{\|D\|_2\leq q} \|D\|_2\|Q\|_2 \text{ (Cauchy-Scwarz)}$$

$$= \sup_{\|D\|_\infty\leq 1} \sup_{\|D\|_2\leq 1} \|D\|_2\|Q\|_2.$$

Since $\|\cdot\|_{170}$ and $\|\cdot\|_2$ are equivalent, the result follows. From Theorem 3.6, it now follows that $\tau' = \tau''$ aare $T_2$-spaces (since $\varepsilon$ is). □

*Theorem 3.8*

$\tau$ is a $T_0$-space, but not a $T_1$-space. Hence $\tau \neq \tau''$.

*Proof.* (i) $\tau$ is $T_0$. Let $D, E \in DS$, $D \neq E$ and we are lookinf for (sufficient condition)

$$[D \in R(D, r) \wedge E \notin R(D, r)]$$

$$\vee$$

$$[D \notin R(E, r) \wedge E \in R(E, r)].$$

Hence, we look for $D = (D_1, \ldots, D_n)$, $E = (E_1, \ldots, E_n)$ in $DS$, such that

$$\left[\langle D, D\rangle = \sum_{i=1}^n D_i^2 > r \wedge \langle D, E\rangle = \sum_{i=1}^n D_iE_i \leq r\right]$$

$$\vee$$

$$\left[\langle E, D\rangle = \sum_{i=1}^n D_iE_i \leq r \wedge \langle E, E\rangle = \sum_{i=1}^n E_i^2 > r\right].$$

It is therefore sufficient to look for $r$, such that

$$\begin{cases} \sum_{i=1}^n D_iE_i \leq r \\ \sqrt{\sum_{i=1}^n D_i^2} \quad \sqrt{\sum_{i=1}^n E_i^2} > r. \end{cases}$$

Such as $r$ exists if

$$\sum_{i=1}^{n} D_i E_i < \sqrt{\sum_{i=1}^{n} D_i^2} \sqrt{\sum_{i=1}^{n} E_i^2}$$

or

$$\langle D, E \rangle < \|D\|_2 \|E\|_2$$

i.e., the inequality of Cauchy-Schwarz must be strict. This is so iff

$$E \notin \{\alpha D \| \alpha \in \mathbb{R}\}.$$

Let now $E = \alpha D$, $\exists \alpha \in \mathbf{R}$. Since $E \in DS$, $\alpha \in \mathbf{R}^+$. If $\alpha > 1$, take $r = \langle D, E \rangle = \alpha \|D\|_2^2$. Then, since

$$\langle E, E \rangle > \langle D, E \rangle$$

we have that $E \in R(E, r)$. Hence, $R(E, r) \in V_\tau(E)$. But $\langle D, E \rangle = r$, hence, $D \notin R(E, r)$. If $\alpha < 1$, then we have the above with $D$ and $E$ replaced: $D \ominus (1/\alpha) E$.

*Proof: (ii) $\tau$ is not $T_1$.* Choose any $D, E \neq 0$ in $DS$ with $E = \alpha D$, $\alpha > 1$. For every $V \in V_{120}(D)$, there is an $n \in 319$, $F_i \in DS$, $r_i > 0$ $(i = 1, \ldots, n)$ such that

$$D \in \bigcap_{i=1}^{n} R(F_i, r_i) \subset V.$$

Hence $\langle F_i, D \rangle > r_i$, $\forall = 1, \ldots, n$ and hence, also $\langle F_i, E \rangle = \alpha \langle F_i, E \rangle > r_i$, $\forall i = 1, \ldots, n$.

Hence,

$$E \in \bigcap_{i=1}^{n} R(F_i, r_i) \subset V.$$

So $(DS, \tau)$ is not a $T_1$-space. Since $\tau' = \tau''$ are (even $T_2$-spaces), we have that $\tau \neq \tau''$. □

**Note.** In the above proof, we had to check carefully if $D \in R(D, r)$. This looks trivial, but it is not so. It is even not always true. Indeed, $D \in R(D, r)$ iff $\langle D, D \rangle > r$, i.e., iff $\|D\|_2 > \sqrt{r}$. So, it can easily be that $R(D, r) \neq \emptyset$, $R(D, r) \in \tau$, but $D \notin R(D, r)$.

## Appendix E: Proof of the Results of Example 3.5.2

*Theorem 3.9*

$$\forall = 1, \ldots, n, \forall r > 0:$$

$R(D_{j,r})$

$$= \left\{ d_i \in DS \middle| i \in \{1, \ldots, n\} \cap \left]2jr - j, \frac{j}{2r-1}\right[ \right\}$$

forming a subbasis for $\tau$.

*Proof.* $\forall i, j \in \{1, \ldots, n\}$, $\forall r > 0$

$$D_i \in R(D_j, r) = \{D \in DS | \text{sim}(D, D_j) > r\} \tag{E1}$$

$\Leftrightarrow$

$$\text{sim}(D_i, D_j) = \frac{i+j}{2 \max(i, j)} > r$$

a) $i < j$

Then Equation E1 is equivalent to

$$j > i > 2jr - j.$$

b) $i > j$

Then, by (a) and reversing $i$ and $j$, we find

$$i > j > 2ir - i.$$

Hence,

$$j < i < \frac{j}{2r-1}. \quad \square$$

*Theorem 3.10*

Let $d$ be the metric of $\tau'$. Then, $\forall i, K = 1, \ldots, n$:

$$d(D_i, D_k) = \frac{|i-k|}{2 \max(i, k)}.$$

*Proof.* $\forall i, k = 1, \ldots, n$, by definition of $d$,

$$d(D_i, D_k) = \sup_{j=1,\ldots,n} \left| \frac{i+j}{2 \max(i, j)} - \frac{k+j}{2 \max(k, j)} \right|.$$

Suppose first that $i \leq k$.

(1) $j \leq i$

Then,

$$\sup_{j=1,\ldots,n} \left| \frac{i+j}{2 \max(i, j)} - \frac{k+j}{2 \max(k, j)} \right|$$

$$\max_{j=1,\ldots,n} \left| \frac{i+j}{2i} - \frac{k+j}{2k} \right| = \frac{k-i}{2k}.$$

(2) $i \leq j \leq k$

Then,

$$\sup_{i\le j\le k}\left|\frac{i+j}{2\max(i,j)}-\frac{k+j}{2\max(k,j)}\right|$$

$$=\max_{i\le j\le k}\left|\frac{i+j}{2i}-\frac{k+j}{2k}\right|$$

$$=\max_{i\le j\le k}\frac{|ik-j^2|}{2kj}=\frac{k-i}{2k}$$

the maximum being obtained in $j = 1$, as well as in $j = k$.
(3) $j \le k$

$$\sup_{j=k,\ldots,n}\left|\frac{i+j}{2\max(i,j)}-\frac{k+j}{2\max(k,j)}\right|$$

$$=\max_{j=k,\ldots,n}\left|\frac{i+j}{2j}-\frac{k+j}{2j}\right|=\frac{k-i}{2k}.$$

So, if $i \le k$,

$$d(D_i, D_k)=\frac{k-i}{2k}.$$

For $i \ge k$ we, hence, have (since $d(D_i, D_k) = d(D_k, D_i)$)

$$d(D_i, D_k)=\frac{i-k}{2i}.$$

This proves the theorem. □

*Corollary 3.11*

$\tau = \tau'' = \tau' = \mathcal{P}(DS)$.

*Proof.* If suffices, by Theorem 2.5, to show that $\tau = \mathcal{P}(DS)$. Now $\forall j = 1, \ldots, n$ and

$$r \in \left]\frac{j+\frac{1}{2}}{j+1}, 1\right[ \neq \varnothing$$

we have that $R(Q, r) = \{Q\}$. This follows readily from Theorem 3.9. □

Note that also $\tau'$ is the discrete topology, but that $d$ is not the discrete metric.

## Appendix F: Proof of the Results of Example 3.5.3

*Theorem 3.12*

$\forall j = 1, \ldots, n, \forall r > 0$

$$R(D_j, r)=\left\{d_i \in DS \mid i \in \{1,\ldots,n\}\right.$$

$$\left.\cap \left]j-\frac{1}{\tan\frac{\pi r}{2}},\, j+\frac{1}{\tan\frac{\pi r}{2}}\right[\ \right\}$$

which forms a subbasis for $\tau$.

*Proof.* $\forall j = 1, \ldots, n$

$$D_i \in R(D_j, r) \Leftrightarrow \frac{2}{\pi}\operatorname{Arctan}\left(\frac{1}{|i-j|}\right) > r$$

(a) $i < j$
Then,

$$D_i \in R(D_j, r) \Leftrightarrow \tan\frac{\pi r}{2} < \frac{1}{j-1}$$

$$\Leftrightarrow j-\frac{1}{\tan\frac{\pi r}{2}} < i < j.$$

An analogous argument yields the right-hand side of the open interval in Equation 24, in case $j < i$. □

*Theorem 3.13*

Let $d$ be the metric of $\tau'$. Then, $\forall i, k = 1, \ldots, n$:

$$d(D_i, D_k)=\frac{2}{\pi}\operatorname{Arctan}|i-k|.$$

*Proof.* $\forall i, k = 1, \ldots, n$:

$$d(D_i, D_k)=\max_{j=1,\ldots,n}\frac{2}{\pi}\left|\operatorname{Arctan}\left(\frac{1}{|i-j|}\right)\right.$$

$$\left.-\operatorname{Arctan}\left(\frac{1}{|j-k|}\right)\right|$$

$$=\frac{2}{\pi}\operatorname{Arctan}\max\tan\left(\left|\left(\frac{1}{|i-j|}\right)\right.\right.$$

$$\left.\left.-\operatorname{Arctan}\left(\frac{1}{|j-k|}\right)\right|\right),$$

since Arctan is an increasing function.

By the two lemmas below, we have that $\forall i, j, k = 1, \ldots, n$

$$\tan\left(\left|\left(\frac{1}{|i-j|}\right) - \operatorname{Arctan}\left(\frac{1}{|j-k|}\right)\right|\right)$$

$$= \left|\tan\left(\operatorname{Arctan}\left(\frac{1}{|i-j|}\right) - \operatorname{Arctan}\left(\frac{1}{|j-k|}\right)\right)\right|$$

$$= \left|\frac{|j-k| - |i-j|}{|i-j||j-k| + 1}\right|$$

using the formula

$$\tan(\alpha - \beta) = \frac{\tan\alpha - \tan\beta}{1 + \tan\alpha\tan\beta}.$$

If we split up the cases, as in the previous example, we find that, in all cases that $i \leq k$, we have that

$$d(D_i, D_k) = \frac{2}{\pi}\operatorname{Arctan}(k - i).$$

The analogue result is true if $i \geq k$, yielding the proof of the theorem. □

*Lemma A*

$$|\tan\alpha| = \tan|\alpha| \text{ if } \alpha \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[.$$

*Lemma B*

$$\operatorname{Arctan}\left(\frac{1}{|i-j|}\right) - \operatorname{Arctan}\left(\frac{1}{|j-k|}\right) \in \left]-\frac{\pi}{2}, \frac{\pi}{2}\right[.$$

*Proof.* For $i, j, k \in \{1, \ldots, n\}$,

$$|i-j|, |j-k| \in [0, n-1].$$

From this, the conclusion follows easily. □

*Corollary 3.14*

$$\tau = \tau'' = \tau = \mathcal{P}(DS)$$

*Proof.*

$$R(D_j, r) = \{D_j\} \text{ if } r > \tfrac{1}{2}:$$

indeed, then

$$j - 1 < j - \frac{1}{\tan\frac{\pi r}{2}}$$

$$j + 1 > j + \frac{1}{\tan\frac{\pi r}{2}}.$$

Hence, $\tau = \mathcal{P}(DS)$ and, hence, also $\tau' = \tau'' = \mathcal{P}(DS)$ *by* Theorem 2.5. □

## References


Cater, S. C. (1986). *The topological information retrieval system and the topological paradigm: A unification of the major models of information retrieval.* Unpublished doctoral dissertation, Louisiana State University, Baton Rouge.

Császár, A. (1978). *General topology.* Disquisitiones Mathematicae Hungaricae 9. Budapest: Akadémiai Kiadó.

Dugundji, J. (1966). *Topology.* Boston: Allyn and Bacon.

Egghe, L., & Rousseau, R. (1997a). Duality in information retrieval and the hypergeometric distribution. *Journal of Documentation, 53,* 488–496.

Egghe, L., & Rousseau, R. (1997b). Everett and Cater's retrieval topology [Letter to the editor]. *Journal of the American Society for Information Science, 48,* 479–481.

Everett, D. M., & Cater, S. C. (1992). Topology of document retrieval systems. *Journal of the American Society for Information Science, 43,* 658–673.

Kreyszig, E. (1978). *Introductory functional analysis with applications.* New York: Wiley.

Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.

Steen, L. A., & Seebach, J. A., Jr. (1978). *Counterexamples in topology.* Berlin: Springer Verlag.

Wilansky, A. (1970). *Topology for analysis.* Lexington, MA: Xerox College.

Willard, S. (1970). *General topology.* Reading, MA: Addison-Wesley.

# On the Necessity of Term Dependence in a Query Space for Weighted Retrieval

**Peter Bollmann-Sdorra**
*Fachbereich Informatik, FR 6-9, Technische Universitat Berlin, Franklin Strasse 28/29, 10587 Berlin, Germany*

**Vijay V. Raghavan**
*Center for Advanced Computer Studies, P. O. Box 44330, University of Southwestern Louisiana, Lafayette, LA 70504-4330. E-mail: raghavan@cacs.usl.edu*

**In recent years, in the context of the vector space model, the view, held by many researchers, that documents, queries, terms, etc., are all elements of a common space has been challenged (Bollmann-Sdorra & Raghavan, 1993). In particular, it was noted that term independence has to be investigated in the context of user preferences and it was shown, through counterexamples, that term independence can hold in the document space, but not in the query space and vice versa. In this article, we continue the investigation of query and document spaces with respect to the property of term independence. We prove, under realistic assumptions, that requiring term independence to hold in the query space is inconsistent with the goal of achieving better performance by means of weighted retrieval. The result that term independence in the query space is undesirable is obtained without making any assumption about whether or not the property of term independence holds in the document space. The results of this article reinforce our position that the properties of document and query spaces must be investigated separately, since the document and query spaces do not necessarily have the same properties.**

## 1. Introduction

The vector space model has been introduced by Salton's SMART group about 30 years ago. Since then, it is one of the most popular models in the sense that most investigations use this model in one way or another. Its popularity comes from the fact that queries and documents are represented both in the same vector notation, which suggests a geometrical interpretation of the retrieval process in an $n$-dimensional space. It suggests the use of many statistical techniques, such as association measures that were developed in numerical taxonomy, and it supports various feedback techniques such as Rocchio's query feedback (Rocchio & Salton, 1965), or more perception-like feedback as in Bollmann and Wong (1987) and Wong, Yao, and Bollmann (1988). In the last decade, there have also been discussions about the underlying assumptions of this model. For example, in Raghavan and Wong (1986), the question was raised whether "vector space model" is just related to vector or tuple notation, or if really a Euclidean vector space was intended. Then, under the assumption that we really have a Euclidean vector space, consequences for the retrieval process were derived. The underlying assumptions of the vector space model were as follows: Documents, terms, and queries are all elements of the same vector space. The term vectors are "independent" in the sense that they constitute an orthogonal basis of the space. It was, of course, accepted that the assumptions of this model might not be fulfilled completely in reality (which is obviously true for many mathematical models that want to capture the reality of natural language processing), but whether these assumptions are consistent was not explored.

This view was challenged recently by Bollmann-Sdorra and Raghavan (1993) by showing that documents and queries behave very differently. To this end, preference relations and evaluation measures, which imply preferences on queries, were exploited. A notion of term independence was introduced, and by means of counterexamples, the possibility of having term independence in the document space, but not in the query space and vice versa, was established. This, in turn, implied that terms are not independent per se, but that term independence has to be investigated in the context of user preferences.

In that article, the investigation was based on examples using unweighted document and query descriptions. In contrast, it is widely agreed that weighted query terms have a potential for better performance. Also, the following question remained open: Would it not be desirable, at least for

Received January 5, 1996; revised October 24, 1997; accepted October 24, 1997.

special cases, that term independence holds both in query and document spaces? In this article, we will show that, for retrieval functions such as dot product or cosine, weighted retrieval is incompatible with term independence in the query space. In other words, in order to capture the benefits of query term weights, term independence is an undesirable property. This result highlights yet another important difference between queries and documents.

Furthermore, we want to mention that this investigation is not meant to discredit the vector processing model as a whole. On the contrary, we ascribe more power to it by obviating the need for certain common restrictions. Our suggestion is that the assumptions of the model (Euclidean vector space, terms are independent, documents and queries are elements of the same space, etc.) are too strong and are not really needed. Finally, this work brings out the importance of investigating properties of a query space separately, since it does not necessarily have the same properties as the associated document space.

## 2. Basic Concepts

### *2.1. The Vector Space Model*

The mathematical model for a typical information retrieval system $S$ can be defined as a 5-tuple $S = (T, D, Q, V, F)$, where: $T$ is the set of index terms; $D$ is the set of documents; $Q$ is a set of queries; $V$ is a subset of the real numbers; $F : D \times Q \rightarrow V$ is a retrieval function between a query and a document (Bookstein & Cooper, 1976). Information Retrieval (IR) systems based on the vector processing model, represent each document by means of index terms (also known as keywords or index words). They are extracted from either the body of a text or a surrogate text (e.g., abstract) through a process of indexing. Documents are represented by a vector of the form $d = (d_1, d_2, \ldots, d_n)$, where $d_i$ is the weight of $t_{di} \in T$. At times, the $d_i$s are binary; however, it is also conventional to assign a weight for each term to reflect the relative importance of the term within a document. A user reveals his/her need either writing a passage in natural language or using keywords connected with Boolean operators. In the former case, the query may be represented within the IR system as a set of (index term, weight) pairs (Salton, 1989). In either case, a query can be expressed in a vector of the form $q = (q_1, q_2, \ldots, q_n)$, where $q_i$ is the weight of $t_{qi} \in T$ (Wong, Ziarko, Raghavan, & Wong, 1989). The result of applying a retrieval function, $F$, is called retrieval status value of $d$ for $q$. Two commonly used retrieval functions are dot product, which is

$$F(q, d) = \sum_{i=1}^{n} q_i d_i$$

and the cosine measure, given by

$$F(q, d) = \frac{\sum_{i=1}^{n}}{\sqrt{(\sum_{i=1}^{n} q_i^2)(\sum_{i=1}^{n} d_i^2)}}$$

The ordered list of documents returned by the IR system to answer a user's query, which is called *the retrieval output,* is obtained by ranking documents with respect to their retrieval status values in descending order.

### *2.2. Document Space and Query Space*

By a document space, we mean a relations system $(D, \cdot \geq)$, where $D$ is a set of document descriptions and $\cdot \geq$ is a preference relation on $D$ (Bollmann & Wong, 1987; Wong et al., 1988). A document description stands for a nonempty equivalence class of documents—for all items with the same description. The preference relation $\cdot \geq$ is defined relative to the information need of some user. We assume that, for the preference relation, the following two conditions hold for all $d, d', d'' \in D$:

i) $d \cdot \geq d'$ or $d' \cdot \geq d$ (connectedness)
ii) $(d \cdot \geq d'$ and $d' \cdot \geq d'') \Rightarrow d \cdot \geq d''$ (transitivity).

In other words, we consider $\cdot \geq$ to be a weak order and $d \cdot \geq'$ is interpreted as: $d$ is preferred to $d'$, or they are equally good. If $d \cdot \geq d'$ and $d' \cdot \geq d$ hold, we write $d \sim d'$, which means that $d$ and $d'$ are equally good. If $d \cdot \geq d'$ and not $d' \cdot \sim d$ holds, we write $d \cdot > d'$, which means that $d$ is better than $d'$. The following is an example of a preference relation:

$d \cdot \geq d'$ iff all documents with description on $d$ are relevant and all documents with description $d'$ are nonrelevant;
$d \sim d'$ both $d$ and $d'$ represent either only relevant documents or only nonrelevant documents.

In the case that there are relevant as well as nonrelevant documents with the same description, we could have the following preference, which is assumed in probabilistic retrieval models: $d \cdot \geq d'$ iff $P$ (relevant | $d$) $\geq P$ (relevant | $d'$).

*Example 2.1.* Let $D \subseteq \{0,1\}^3$. Suppose that the preference on the document space $(D, \cdot >)$ is given by

$$\left\{\begin{matrix} (1, & 0, & 0) \\ (1, & 0, & 1) \\ (1, & 1, & 0) \end{matrix}\right\} \cdot > \left\{\begin{matrix} (0, & 0, & 1) \\ (0, & 1, & 0) \\ (0, & 1, & 1) \\ (1, & 1, & 1) \end{matrix}\right\}.$$

Documents in { }are considered to be equally good for the user. Three of them, (1, 0, 0), (1, 0, 1), and (1, 1, 0), are relevant, whereas the remaining four are nonrelevant. #

A query space is a relational system denoted by $(Q, \cdot \geq^*)$, where $Q$ is the set of all possible queries and $\cdot \geq^*$ is a preference relation on $Q$. Preference relations on $Q$ were first considered implicitly by Rocchio and Salton (1965) in

the sense that the generation of an optimal query always involves some preference. Preference relations on $Q$ were also considered by Yu (1974), Yu and Salton (1976), and Robertson, Maron, and Cooper (1982). The works of Rocchio, Salton, and Yu relate more closely to this investigation, since they also considered preference on queries with respect to a particular user need. In Yu, Salton, and Siu (1978), for example, it is shown that a study of query space has implications for query modifications through term addition and deletion. Robertson et al. (1982), on the other hand, consider a family of queries, where each instance represents a different user need.

The preference relation on $Q$ depends on the following factors: The document space $(D, \cdot \geq)$, the retrieval function $F: D \times Q \rightarrow R$, which maps each document-query pair into the real numbers, and the evaluation measure that assesses the rankings generated by $F$. We illustrate this in the following example. In the example, the evaluation measure $R_{norm}$, defined in the Appendix, is chosen. $R_{norm}$ is intended to compare the rank ordering of documents generated by the system, relative to the preference ordering of the same documents given by the user, with respect to a given need. $R_{norm}$ reaches the highest value of 1, if, for all $d, d' \in D$, $d \cdot > d'$ implies that system ranks $d$ higher than $d'$.

*Example 2.2.* We consider the document space of Example 2.1. Let $Q = D$. The retrieval function $F$ is given by the coefficient of Jaccard, where

$$F(d, q) = \frac{\text{\# of terms in } d \text{ and } q}{\text{\# of terms in } d \text{ or } q}$$

The evaluation measure is the normalized recall, $R_{norm}$. Let $q = (1, 1, 0)$. This gives the following similarities for the documents:

$$F((0, 0, 1), q) = 0$$

$$F((0, 1, 0), q) = 5$$

$$F((0, 1, 1), q) = .333$$

$$F((1, 0, 0), q) = .5$$

$$F((1, 0, 1), q) = .333$$

$$F((1, 1, 0), q) = 1.0$$

$$F((1, 1, 1), q) = .667$$

The resulting ranking of the documents is as follows:

$$\left((1, 1, 0) \mid (1, 1, 1) \left| \begin{matrix}(0, 1, 0)\\(1, 0, 0)\end{matrix} \right| \begin{matrix}(0, 1, 1)\\(1, 0, 1)\end{matrix} \right| (0, 0, 1)\Big)$$

If we substitute each document by its relevance judgments, we obtain the ranking

$$( + \mid - \mid \pm \mid \pm \mid - )$$

where the $+$ signs and $-$ signs represent the rank of, respectively, the relevant and nonrelevant documents.

For this ranking, which corresponds to setting $q = (1, 1, 0)$, we obtain $R_{norm} = \frac{16}{24} = .667$. If we repeat this for all queries, we obtain the following preference relation on queries, defined by their $R_{norm}$ values:

| | $R_{\text{norm}}$: |
|---|---|
| $(1, 0, 0)$ | 1 |
| $\cdot >^* \left\{\begin{matrix}(1, 0, 1)\\(1, 1, 0)\end{matrix}\right.$ | .667 |
| $\cdot >^*(1, 1, 1)$ | .583 |
| $\cdot >^* \left\{\begin{matrix}(0, 0, 1)\\(0, 1, 0)\end{matrix}\right\}$ | .292 |
| $\cdot >^*(0, 1, 1$ | 0 |
| | # |

## 2.3. Term Independence

We have already seen that term independence is considered to be one of the basic assumptions of the vector space model. If term independence is viewed as an empirical property, it should have a definition such that the conduct of experiments in order to find out whether such an assumption holds in reality is possible. A notion of term independence was introduced in the context of probabilistic retrieval models. There, term independence means that terms contribute, independently of each other, to the probability that a document with certain description is relevant. Hence, the general principle that we want to adopt is that term independence means that terms contribute independently of each other to some well-defined property. This means, in order to discuss term independence among terms in queries, we need an analogous property. Our review of IR literature shows that the property that is mostly investigated for queries is how well queries perform with respect to a given retrieval function and an evaluation measure (Robertson & Sparck-Jones, 1976; Robertson et al., 1982; Rocchio & Salton, 1965; Wong et al., 1988; Yu & Salton, 1976). This is also supported by the fact that all query feedback methods are intended to find some optimal query. Hence, we think, for information retrieval, that it is natural to discuss term independence in the query space in the context of preferences on queries. For subsequent developments, we require a definition of term independence that is applicable to both document and query space. The probabilistic concept of term independence that is used for the document space, which is based on a frequency ration interpretation of probability,

does not easily generalize to the query space. We, therefore, consider a notion of independence that has been introduced in measurement theory (Krantz, Luce, Suppes, & Tversky, 1971, chap. 7). It is defined as follows:

*Definition 2.1.* Let $A = A_1 \times \cdots \times A_n$ be the Cartesian product of $n$ nonempty sets. Let $\cdot \geq$ be a binary relation on $A$. $A_i$ is *independent* of all other attributes $A_j$ (that is, $j \neq i$), iff the following condition holds for all $i$, $1 \leq i \leq n$, $a_i$, $a_i' \in A_i$, and for all $b = (b_1, \ldots, b_n)$, $c = (c_1, \ldots, c_n)$ where $b$, $c \in A$:

$$
\begin{aligned}
&(b_1, \ldots, b_{i-1}, a_i, b_{i+1}, \ldots, b_n) \\
&\qquad \cdot \geq (b_1, \ldots, b_{i-1}, a_i', b_{i+1}, \ldots, b_n) \\
&\qquad \Leftrightarrow (c_1, \ldots, c_{i-1}, a_i, c_{i+1}, \ldots, c_n) \\
&\qquad\qquad \cdot \geq (c_1, \ldots, c_{i-1}, a_i', c_{i+1}, \ldots, c_n) \;\#
\end{aligned}
$$

This means that, for $A_i$ to be independent of all other attributes, if the substitution of $a_i'$ by $a_i$ gives an improvement in one context, then it should give an improvement in every context.

Sometimes, in information retrieval, not all objects of the Cartesian product are available. For example, if the retrieval function is the cosine measure, then the zero vector is not admissible. In that case, we assume the condition of the definition above to hold only for the admissible vectors.

In the following example, we show that it is possible that term independence holds in the document space, but not in the query space.

*Example 2.3.* Let $D = Q = \{1,0\}^3$. We consider the following preference relation on $D$.

$$
\begin{Bmatrix} (1,1,0) \\ (1,1,1) \end{Bmatrix} \cdot > \begin{Bmatrix} (0,1,0) \\ (0,1,1) \\ (1,0,0) \\ (1,0,1) \end{Bmatrix} \cdot > \begin{Bmatrix} (0,0,0) \\ (0,0,1) \end{Bmatrix}
$$

Here the terms are independent. For term $t_1$ we obtain:

$$(1, 0, 0) \cdot > (0, 0, 0)$$

$$(1, 0, 1) \cdot > (0, 0, 1)$$

$$(1, 1, 0) \cdot > (0, 1, 0)$$

$$(1, 1, 1) \cdot > (0, 1, 1)$$

Hence, we see that the inclusion of term $t_1$ always gives an improvement, no matter what the other term values are. Hence, term $t_1$ is independent of the other terms. Similar considerations can be also made for the terms $t_2$ and $t_3$.

We now choose dot product as the retrieval function, and the normalized recall, $R_{norm}$, as the evaluation measure. This gives the following query space $(Q, \cdot \geq^*)$:

| | $R_{\text{norm}}$: |
|---|---|
| $(1, 1, 0)$ | 1 |
| $\cdot >^* (1, 1, 1)$ | .9 |
| $\cdot >^* \begin{Bmatrix} (1,0,0) \\ (0,1,0) \end{Bmatrix}$ | .8 |
| $\cdot >^* \begin{Bmatrix} (1,0,1) \\ (0,1,1) \end{Bmatrix}$ | .725 |
| $\cdot >^* \begin{Bmatrix} (0,0,1) \\ (0,0,0) \end{Bmatrix}$ | .5 |

Here, term independence does not hold in the query space, as evidence by the fact that the third term is not independent of the first two; that is, $(1, 1, 0) \cdot >^* (1, 1, 1)$, but $(0, 0, 0) \sim^* (0, 0, 1)$. #

Whenever independence holds for document vectors, according to Definition 2.1, we see that terms contribute to document quality independently of each other; that is, it is an essential property for a linear decision function to exist.

In the next section, we show that in the case that the query terms are weighted, insistence on term independence in the query space can negate the effect of weights.

## 3. Weighted Retrieval and Term Independence in the Query Space

It is widely accepted that weighted query terms offer the potential for improving retrieval performance. Especially, all query feedback methods are aimed at deriving the term weights for the optimal query. In the following theorems, we show that for a certain class of retrieval functions, which may include either one or both dot product and cosine measures, weighted retrieval is incompatible with term independence in the query space. To this end, we first closely consider the various assumptions needed.

One of the assumptions in Theorem 3.1 is that the document space is finite. Although, in many retrieval models, there are infinitely many document descriptions (as in the vector space model with weighted document terms), a user will always have to make retrieval on a finite document collection. This implies that the set of document descriptions under consideration at any given time will be finite. Hence, we think it is quite natural to assume that the document space is finite.

There are basically two ways to evaluate retrieval results. One approach is based solely on the ranking of the documents and it disregards the retrieval status values. Measures that are based on this view are exemplified by recall-precision graph, recall-fallout graph, normalized recall, normalized precision, and expected search length. They constitute a majority of the evaluation measures so far designed. Cherniavsky and Lakhuty (1970) call this evaluation *relative* because only the positions of the documents relative to each other is taken into consideration. The alternative approach of evaluation uses the retrieval status values. Cher-

niavsky and Lakhuty called such an evaluation *absolute.* Examples of absolute measures are the measure of Swets (1969), or the measure used in the SMART project in order to define the optimal query (Rocchio & Salton, 1965). We note that, with respect to query term independence, whether the evaluation is relative or absolute can make a difference, since we want to make use of the property that the preference relation on the query space can only have finitely many levels. In Theorem 3.1, we assume that relative evaluation is used. In general, in IR, there are sound theoretical reasons to favor relative evaluation measures (Robertson, 1977).

In the vector space model, among the most important retrieval functions are the dot product and the cosine measure. Both retrieval functions have a common property. If we multiply a query vector $q = (q_1, \ldots, q_n)$ with some constant $\lambda > 0$, then $\lambda q = (\lambda q_1, \ldots, \lambda q_n)$ gives the same ranking of documents as $q$, for both functions. In the case of dot product, we have

$$F(q, d) \geq F(q, d') \Leftrightarrow F(\lambda q, d) \geq F(\lambda q, d')$$

Hence, we obtain exactly the same ranking of document. In the case of the cosine measure, we obtain

$$F(q, d) = F(\lambda q, d)$$

which also implies the same ranking. In the case of a relative evaluation, in both cases $\lambda q \sim * \, q$ holds. For both Theorems 3.1 and 3.2, we shall, therefore, assume a class of retrieval functions with the property $q \sim * \, \lambda q$. We think that this class of retrieval functions is of special interest in information retrieval, since it includes both dot product and cosine. It should be noted that for some absolute evaluation measures (e.g., measure of Swets), the property $q \sim * \, \lambda q$ holds for dot product.

The following theorem states that under certain assumptions three query weights, $-1$, 0 and $+1$, are sufficient. By this we mean, that for each query $q$ there exists a query $q'$, such that all the term weights of $q'$ are one of $-1$, 0, or $+1$ and $q \sim * \, q'$.

*Lemma 3.1.* Assume that there exist a query $q'_0 = (q'_1, \ldots, q'_n)$ and a query $q''_0 = (q''_1, \ldots, q''_n)$, such that

$$q''_i = \begin{cases} 1 & q'_i > 0 \\ 0 & q'_i = 0 \\ -1 & q'_i < 0 \end{cases}$$

and $q'_0 \not\sim * \, q''_0$. Then, there exist two queries, $q' = (q'_1, q^*_2, \ldots, q^*_n)$ and $q'' = (q''_1, q^*_2, \ldots, q^*_n)$, such that $q'_1/q''_1 > 0$ and $q' \cdot > * \, q''$.

**Proof.** Let $q''_k = (q''_1, \ldots, q''_k, q'_{k+1}, \ldots, q'_n)$, for $k = 0, 1, \ldots, n$. Then $q'_{00} = q'_0$ and $q'_{0n} = q''_0$. If, for all $k$, $q'_{0k} \sim^* q'_{0(k+1)}$ then, because of transitivity $q'_0 \sim^* q''_0$. Hence, there exist two queries, $q'_{0k} = (q''_1, \ldots, q''_k, q'_{k+1}, \ldots, q'_n)$ and $q'_{0(k+1)} = (q''_1, \ldots, q''_k, q''_k, q''_{k+1}, q'_{k+2, \ldots,} q'_n)$, that are not equivalent. They only differ in the term $k + 1$, and also $q'_{k+1}$ and $q''_{k+1}$ have the same sign. By rearrangement of the terms, we may assume, without loss of generality, that they only differ in the first term.

*Theorem 3.1.* Assume that the following assumptions hold:

i) The document space if finite;
ii) $Q, \cdot \geq^*$) is a weak order, and the preference relation is based only on the ranking of the documents as defined by the query and the retrieval function, (i.e., the evaluation measure is relative);
iii) For all $q \in Q$ and $\lambda > 0$, and $\lambda \cdot \; q \in Q$ and, $q \sim^* \lambda q$; and
iv) Term independence holds in the query space.

Then for each query term the three weights, $-1$, 0, and $+1$, are sufficient.

**Proof.** The proof will be indirect. Assume that there exists a query term for which the three weights $-1$, 0, $+1$ are not sufficient. Without loss of generality, we assume it to be the first term. Hence, because of Lemma 3.1, there exist two queries, $q'$ and $q''$ with $q' \cdot >^* q''$, $q' = (q'_1, q^*_2, \ldots, q^*_n)$ and $q'' = (q''_1, q^*_2, \ldots, q^*_n)$, such that $q'$ and $q''_1$ are both non-zero and have the same sign. Let $\lambda = q'_1/q''_1 > 0$.

From this we obtain, because of assumption iii,

$$\begin{aligned}(\lambda^{k+1} q''_1, \lambda^k q^*_2, \ldots, \lambda^k q^*_n) &\sim *(\lambda q''_1, q^*_2, \ldots, q^*_n) \\ &= (q'_1, q^*_2, \ldots, q^*_n) \cdot > *(q''_1, q^*_2, \ldots, q^*_n) \\ &\sim (\lambda^k, \lambda^k q^*_2, \ldots, \lambda^k q^*_n)\end{aligned}$$

Because of the independence assumption iv, this implies $(\lambda^{k+1)} q''_1, q_2, \ldots, q_n) \cdot >^* (\lambda^k q''_1, q_2, \ldots, q_n)$ *for all* $q_i \in R$, $i = 2, \ldots, n$. This then implies that there are infinitelymany preference levels in $(Q, \cdot \geq^*)$. On the other hand, the finiteness of the document space (assumption i) and the adoption of relative evaluation measure (assumption ii) imply that there are only finitely many rankings of the document space. Since, by assumption ii), only these rankings are evaluated, $(Q, \cdot \geq^*)$ can have only finitely many preference levels. Hence, we have a contradiction. Let $q = (q_1, \ldots, q_n) \in Q$ be an arbitrary query. We now define $q' = (q'_1, \ldots, q'_n)$ as

$$q'_i = \begin{cases} 1, & \text{if } q_i > 0 \\ 0, & \text{if } q_i = 0 \\ -1 & \text{if } q_i < 0 \end{cases}$$

Then $q \sim^* q'$. Hence, the three weights $-1$, 0, and $+1$ are sufficient. #

The theorem can be used to derive several corollaries. For example, in a real retrieval situation, the three weights $-1$, 0, and $+1$ will not be sufficient. Hence, we can conclude that under the given assumptions, term independence will not hold in the query space. Another conclusion is that

if term independence holds, and for some reason negative query weights are not included, then unweighted queries perform as well as weighted ones. This basically means that, under the given assumptions, term independence in the query space holds only for unweighted queries. Moreover, it shows that, for the query space, term independence is not a desirable property. It also should be noted that the only assumption made about the document space is that it is finite. There are no assumptions made as to whether the document terms are weighted, or whether term independence holds in the document space. This again shows that documents and queries are fundamentally different objects, and that they should not be treated in a uniform manner.

We next present a theorem that deals with the incompatibility between term independence in the query space and weighted retrieval in a more general setting. Before presenting the theorem, we define the notion of an optimal query.

*Definition 3.1.* Let $(Q, \cdot \geq^*)$ be a query space. then $q^* \in Q$ is an optimal query iff $q^* \cdot \geq^* q$ for all $q \in Q$.

*Lemma 3.2.* Assume that there exist an optimal query $q^* = (q_1^*, \ldots, q_n^*)$ and a query $q' = (q_1', \ldots, q_n')$, such that

$$q_i' = \begin{cases} 1 & q_i^* > 0 \\ 0 & q_i^* = 0 \\ -1 & q_i^* < 0 \end{cases}$$

and not $q^* \sim^* q'$. Then there exist an optimal query $q_0^* = (q_{01}^*, \ldots, q_{0n}^*)$ and a query $q_0' = (q_{01}', q_{02}^*, \ldots, q_{0n}^*)$, such that $q_{01}^*/q_{01}' > 0$ and $q_0^* \cdot >^* q_0'$.

**Proof.** Similar to the proof of Lemma 3.1.

*Theorem 3.2.* Assume the following holds:

ii) There exists an optimal query $q^* = (q_1^*, \ldots, q_n^*)$.
ii) $(Q, \cdot \geq^*)$ is a weak order. iii) For all $q \in Q$ and $\lambda > 0$, $\lambda \cdot q \in Q$ and $q \sim^* \lambda q$.
iv) Terms are independent in the query space.

Then there exists an optimal query $q' = (q_1', \ldots, q_n')$, such that $q_i' \in$ for all $i = 1, \ldots, n$ and $q' \sim^* q^*$.

**Proof.** The proof is similar to the proof of Theorem 3.1. Assume such a $q'$ does not exist. Then, because of Lemma 3.2, we can conclude that there exists a query $q = (q_1, q_2^*, \ldots, q_n^*)$, such that

$$(q_1^*, {}_2^*, \ldots, q_n^*) \cdot >^* (q_1, q_2^*, \ldots, q_n^*)$$

and $q_1^*/q_1 = \lambda > 0$.

As in the proof of Theorem 3.1, we conclude

$$(\lambda^{k+1} q_1, \lambda^k q_2^*, \ldots, \lambda^k q_n^*) \cdot >^* (\lambda^k q_1, \lambda^k q_2^*, \ldots, \lambda^k q_n^*), \ \forall\, k = 1, 2, \cdots$$

and because of term independence in the query space

$$(\lambda^{k+1} q_1, q_2^*, \ldots, q_n^*) \cdot >^* (\lambda^k p_1, q_2^*, \ldots, q_n^*)$$

For $k = 1$, we obtain

$$\lambda q_1^*, q_2^*, \ldots, q_n^*) \cdot >^* (q_1^*, q_2^*, \ldots, q_n^*)$$

But this is a contradiction because $(q_1^*, \ldots, q_n^*)$ is an optimal query. #

In the theorem above, there are no assumptions made about the finiteness of the document space. Another important difference to note between Theorems 3.1 and 3.2 is that the latter does not mention the type of evaluation measure. If the evaluation measure is absolute, it can happen that $q$ and $\lambda q$ produce the same ranking, but $q \not\sim^* \lambda q$. In such a case, Theorem 3.2 holds only for cosine retrieval function. In contrast, if relative evaluation is used, then Theorem 3.2 applies for both dot product and cosine. Of course, because of fewer assumptions, conclusions are a bit weaker. We cannot say that three weights
are sufficient for any query; rather, that result holds only for optimal queries. But this does not make the situation any better. Why should we use weights if, for optimal queries, the weights $\{-1, 0, +1\}$ are sufficient for any query; rather, that result holds only for optimal queries. But this does not make the situation any better. Why should we use weights if, for optimal queries, the weights $\{-1, 0, +1\}$ are sufficient? Hence, again, term independence in the query space contradicts the benefits of weights. Of course there is always the possibility that, in the general case, no optimal query exists. But in any case, IR in such a context is a hopeless endeavor.

The notion of independence that is involved in Definition 2.1 is a kind of utility theoretic independence. In the case of the probabilistic binary independence model, there exists a linear decision function in the document space. Probabilistic independence implies utility theoretic independence in the document space. But from Theorem 3.2, we can conclude that, even under the assumptions of the probabilistic binary independence model, query term independence does not hold.

In the following example, we show a particular situation where term independence holds in the query space and the weights $-1$, 0, and $+1$ are not sufficient.

*Example 3.1* Let $Q \subseteq R^n$ and let the retrieval function be $F(q, d) = \sum_{i=1}^n f_i\ (q_i, d_i)$. Examples of such a retrieval function are dot product and Canberra metric (Sneath & Sokal, 1973). Furthermore, let the evaluation measure, $\mu$, be $\mu(q = \alpha \sum_{d \in R} F(q, d) - \beta \sum_{d \in \bar{R}} F(q, d)$, where $\alpha, \beta > 0$, $R$ is the set of relevant documents and $\bar{R}$ is the set of nonrelevant documents. Such an evaluation was, for example, chosen in the SMART-project in the context of query feedback. Usually in such a case, three query term weights are not sufficient. We now obtain

$$\mu(q) = \alpha \sum_{d \in R} F(q, d) - \beta \sum_{d \in \bar{R}} F(q, d)$$

$$= \alpha \sum_{d \in R} \sum_{i=1}^{n} f_i(q_i, d_i) - \beta \sum_{d \in \bar{R}} \sum_{i=1}^{n} f_i(q_i, d_i)$$

$$= \sum_{i=1}^{n} \left( \alpha \sum_{d \in R} f_i(q_i, d_i) - \beta \sum_{d \in \bar{R}} f_i(q_i, d_i) \right)$$

If we now define

$$\phi_1(q_i) = \alpha \sum_{d\in R} f_i(q_i, d_i) - \beta \sum_{d\in \bar{R}} f_i(q_i, d_i)$$

then $\mu(q) = \sum_{i=1}^{n} \phi_i (q_i)$.

Then for the preference relation $\cdot \geq^*$ defined as $q \cdot \geq^* q' \Leftrightarrow \mu(q) \geq \mu(q')$, it is easy to see that term independence property holds in the query space. Unfortunately, however, according to Theorem 3.2, there are infinitely many preference levels in $(Q, \cdot \geq^*)$. Thus, if one insists on term independence, optimal query cannot exist and, Rocchio's objective of finding an optimal query is unattainable. #

## 4. Conclusions

In a previous article, we (Bollmann-Sdorra & Raghavan, 1993) showed by a number of examples, that properties of a document space can be very different from those of the associated query space. In contrast, in this article, we formally establish that, with respect to the property of term independence, requiring query and document spaces to be identical has undesirable consequences.

It is proved that, under realistic assumptions, in order to realize the benefits of weighted retrieval, the property of term independence must not hold in the query space. Moreover, term independence in the query space is found to be undesirable, with no assumptions made regarding whether such a property holds in the document space.

What are the consequences of these results for the vector space model? One commonly accepted assumption for the vector space model is that terms are independent, and documents and queries are elements of the same space. Salton himself considered these as disadvantages (Salton, 1989, p. 319). By not requiring the assumption that documents and queries belong to a common space, we have shown that the disadvantages mentioned above do not exist. More specifically, document term independence is a desirable property, but not query term independence. In this sense, our results give the vector space model a greater generality (by removal of assumptions originally thought to be required) and broaden its scope of applicability.

Our work also suggests that more detailed studies that explore the properties of query spaces are needed. For example, in Raghavan and Sever (1995), properties of the query space are investigated form the point of reusing optimal queries.


## Acknowledgments

This work is supported in part by grants from Louisiana Board of Regents Support Fund (under Grant LEQSF (1997–00)-RD-B-16) and the U.S. Department of Energy (under Grant DE-FG02-97ER1220).


## Appendix

*Definition of $R_{norm}$*

Let $(D, \cdot \geq)$ be a document space, where $D$ is a finite set of documents on which a preference relation, $\cdot \geq$, that is complete and transitive (weak order) is defined. Let $\Delta$ be some ranking of $D$ given by a retrieval system. Then $R_{norm}$ is defined as

$$R_{\text{norm}}(\Delta) = \frac{1}{2}\left(1 + \frac{S+ \;-S-}{S^+_{\max}}\right)$$

where $S^+$ is the number of document pairs where a better document is ranked ahead of a worse one, $S^-$ is the number of pairs where a worse document is ranked ahead of a better one, and $S^+{}_{max}$ is the maximal possible number of $S^+$. This $R_{norm}$ was introduced in the LIVE-Project (Bollmann, Jochum, Reiner, Weissmann, & Zuse, 1985) and further investigated by Yao (1995).

*Example.* We consider the ranking

$$\Delta = (rp|pn|rnn|nnn)$$

where $r$ stands for a relevant document, $p$ for a partially relevant document, and $n$ for a nonrelevant document. There are two relevant documents, two partially relevant documents, and six nonrelevant documents. Vertical lines are used as rank separators. That is, first rank has one relevant and partially relevants document, etc. Hence $S^+{}_{max} = 28$, $S^+ = 21$, $S^- = 3$, and $R_{norm}(\Delta) = .82$.

## References


Bollmann, P., Jochum, F., Reiner, U., Weissmann, V., & Zuse, H. (1985). The LIVE project-retrieval experiments based on evaluation viewpoints. In V. Raghavan & P. Gallina (Eds.), *Proceedings of the Eighth Annual International ACM/SIGR Conference on Research and Development in Information Retrieval* (June 1985, Montreal, CDN, pp. 213–214). New York: The Association of Computing Machinery.

Bollmann, P., & Wong, S. K. M. (1987). Adaptive linear information retrieval models. *Proceedings of the Tenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval* (June 1987, New Orleans, LA, pp. 157–163).

Bollmann-Sdorra, P., & Raghavan, V. V. (1993). On the delusiveness of adopting a common space for modeling IR objects: Are queries documents? *Journal of the American Society for Information Science,* 579–587.

Bookstein, A., & Cooper, W. (1976). A general mathematical model for information retrieval systems. *Library Quarterly, 46(2),* 153–167.

Cherniavsky, V. S., & Lakhuty, D. G. (1970). Problem of evaluating retrieval systems I. *Nauchno-Techniceskaya Informazia* (Ser. 2, pp. 24–30) [in Russian]. [English Translation: *Automatic documentation and mathematical linguistics* (Vol. 4, pp. 9–26).]

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement* (Vol. 1). New York: Academic Press.

Raghavan, V. V., & Sever, H. (1995). On the reuse of past optimal queries. In E. A. Fox, P. Ingwersen & R. Fidel (Eds.), *Proceedings of the ACM SIGIR '95* (July 1995, Seattle, WA, pp. 344–350). New York: The Association of Computing Machinery.

Raghavan, V. V., & Wong, S. K. M. (1986). A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science, 33,* 279–287.

Robertson, S. E., (1977). Theories and models in information retrieval. *Journal of Documentation, 33,* 126–149.

Robertson, S. E., Maron, M. E., & Cooper, W. S. (1982). Probability of relevance: A unification of two competing models for document retrieval. *Information Technology: Research and Developoment 1(1,* 1–21.

Robertson, S. E., & Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science, 27,* 129–146.

Rocchio, J. J., & Salton, G. (1965). Information optimization and interactive retrieval techniques. *Proceedings of the AFIPS-Fall Joint Computer Conference, 27*(Pt. I), 293–305.

Salton, G. (1989). *Automatic text processing.* Reading, MA: Addison Wesley.

Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical taxonomy: The principles and practice of numerical classification.* San Francisco: Freeman.

Swets, J. A. (1969). Effectiveness of information retrieval methods. *American Documentation, 20(1),* 72–89.

Wong, S. K. M., Yao, Y. Y., & Bollmann, P. (1988). Linear structure in information retrieval. In Y. Chiaramella (Ed.), *Proceedings of the Eleventh Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval* (June 1988, Grenoble, France, pp. 219–232). Grenoble: Presses Universitaires de Grenoble.

Wong, S. K. M., Ziarko, W., Raghavan, V. V., & Wong, P. C. N. (1989). Extended Boolean query processing in the generalized vector space model. *Information Systems, 14(1),* 47–63.

Yao, Y. Y. (1995). Measuring retrieval effectiveness based on user preferences on documents. *Journal of the American Society for Information Science, 46,* 133–145.

Yu, C. T. (1974). A clustering algorithm based on user queries. *Journal of the American Society for Information Science, 25,* 218–226.

Yu, C. T., & Salton, G. (1976). Precision weighting—An effective automatic indexing method. *Journal of the Association for Computing Machinery, 23,* 76–88.

Yu, C. T., Salton, G., & Siu, M. K. (1978). Effective automatic indexing using term additioin and deletion. *Journal of the Association for Computing Machinery, 25,* 210–225.

# Optimizing a Library's Loan Policy: An Integer Programming Approach

**Hesham K. Al-Fares**
*Systems Engineering Department, King Fahd University of Petroleum and Minerals, P.O. Box 2003, Dhahran 31261, Saudi Arabia. E-mail: hesham@ccse.kfupm.edu.sa*

**Providing the best service in order to satisfy users is the main objective of any library. The loan policy is a major tool available to achieve this objective. Previous studies have all focused on the loan period, ignoring the loan policy's other equally important aspect, which is the maximum number of books a user can borrow. Moreover, only book availability has been used as a measure of user satisfaction. User satisfaction with either the length of the loan period or the number of books allowed to be borrowed has been largely overlooked. This article combines those relevant components into a more complete model. An easy-to-solve integer programming model is formulated, whose solution yields the optimum loan period and optimal limit on the number of books that can be borrowed to maximize user satisfaction. A case study of an actual university library is presented.**

## 1. Introduction

With resources shrinking, and collections and the number of users ever growing, library management is an increasingly difficult task. Hindle and Buckland (1976) maintain that the main objective of a library, as a service facility, is to satisfy its users. They classify library management into the following decision areas: Acquisition, collection control, user education, and administrative systems. The major tool available to library management for achieving its user satisfaction objective is the collection control, or loan policy. While the loan policy has a large influence on user satisfaction, it is also directly under the control of the librarians.

According to Buckland (1975, p. 75), three regulations constitute a loan policy: (1) The maximum loan period, (2) the maximum number of books that can be borrowed by the same user, and (3) the number of renewals allowed. For the sake of completeness, we can add a fourth regulation: One that concerns reservations and recalls. Of these four regulations, the first two are much more important. Buckland (1972) finds that while there is a strong tendency for books to be kept out until their due date, the great majority of loans are not renewed. Somewhat surprisingly, even the length of the official loan period has little effect on the frequency of renewals. Moreover, Buckland (1975, p. 91) also observes that reservations are made only on a small fraction of instances when books are not immediately available.

This article is mainly concerned with the two more important aspects of the loan policy: (1) The length of loan period, and (2) the number of books allowed to be borrowed. The objective is to maximize user satisfaction with both of them, in addition to satisfaction with book availability. The article is organized as follows. First, a review of previous relevant literature is given. Then, the integer programming model of the problem is formulated. Subsequently, considerations for implementing the model are discussed. Next, a case study is presented, and availability percentages are compared with previously published values. Finally, conclusions and recommendations are given.

## 2. Literature Review

The application of operations research (OR) techniques to library problems was pioneered by Morse (1968, 1979). In his book, *Library Effectiveness: A Systems Approach,* he relies primarily on queuing theory and Markovian processes to model circulation and book use. Morse also uses queuing theory to analyze loan policies, in particular the effect of reducing the loan period on satisfying circulation demand. For library effectiveness, the criterion that Morse uses is the ability to provide service (that is, the availability of desired books) to users.

Using computer simulation, significant contributions have been made by the library research group at the University of Lancaster (Buckland, 1972, 1975; Hindle & Buckland, 1976). Their main objective is maximizing book availability, which they consider to be the result of three

Received September 25, 1996; revised August 1, 1997; accepted October 30, 1997.

interacting factors: (1) The frequency of demand, (2) the length of the loan period, and (3) the amount of duplication. Two measures of effectiveness are defined: (1) Satisfaction level: The average immediate availability for a given time duration; and (2) collection bias: The average availability of the 10% most requested books. The group's main recommendation is to implement a variable loan policy, in which the most popular books are subject to a shorter loan period.

Several studies related to loan period, availability, and satisfaction have been conducted at Case Western Reserve University. Kantor (1976b) proposes using document exposure time and user satisfaction level as a combined measure of library performance. Kantor (1976a) uses a branching diagram to describe how availability is reduced by four factors: Lack of acquisition, circulation, misshelving, and user error. Shaw (1976a) develops a simulation model that evaluates the effect of the loan period on user satisfaction, which is measured by book availability and delay associated with recalls. Shaw (1976b) conducts a survey of loan period distribution in academic libraries. Saracevic, Shaw, and Kantor (1977) use a branching approach to analyze the causes of user dissatisfaction, including the loan period.

An approach employing basic queuing theory has been used by Goyal (1970) to determine the optimal loan periods for periodicals. Two models are proposed, both considering readers as customers, periodicals as servers, and the loan period as service time, which is considered exponentially distributed as a simplifying assumption. The objective of the first model is to minimize the total cost of providing the service plus the cost of waiting. The objective of the second model is to maximize customer satisfaction, assumed to be a function of the loan period, book availability, and waiting time.

Bookstein (1975) develops a more sophisticated analytical model based on queuing theory in order to determine optimal loan periods. The model assumes that no queues are allowed to form (no reservations permitted), but it considers the effect of user satisfaction with the loan period. Considering transactions at the circulation desk as generating costs, the objective of the model is to minimize the number of these transactions. Alternative objective functions, such as maximizing book availability or maximizing the probability of successful book use, are also discussed.

Several studies deal with the loan policy and its relationship to various aspects of the library system. Burkhalter and Race (1968) analyze the effect of renewals, overdues, and other factors on the optimal loan period. Bruce (1975) models the library loan dynamics as a four-state Markovian system, relating the loan period to this system. Yagello and Guthrie (1975) examine the increase in circulation resulting from reducing the loan period. Sinha and Clelland (1976) present a general model for collection control, which can be extended to include variations in the loan period. Burrell (1980) proposes, and Hindle and Worthington (1980) discuss, simple stochastic models for library loans which explain some empirical circulation frequency distributions. Burrell and Fenton (1994) develop a modified model of library circulation which accounts for book unavailability for borrowing while it is out on loan. Goehlert (1978) analyzes the influence on availability by several factors, including book reshelving, search, circulation, on order, not owned, duplication, age, and subject. Goehlert determines that circulation is the most important factor. Kantor (1979) presents a comprehensive review of library operations research up to 1979. Kraft and Boyce (1991) provide the most recent and extensive bibliography of library OR literature.

## 3. The Integer Programming Model

The literature review confirms the importance of loan policy for library user satisfaction. The four elements of the loan policy have been identified as: Loan period, number of books allowed, renewals, and reservations or recalls. The first two are shown to be more important; however, all of the reviewed research seems to focus only on the loan period. Although the maximum number of books to be borrowed is also significant for achieving user satisfaction, it has been apparently overlooked in the literature. The model to be presented here considers the two major elements of the loan policy: (1) The length of the loan period, and (2) the maximum number of books allowed to be borrowed.

User satisfaction appears to be a common objective based on all the literature surveyed. For measuring satisfaction, most articles simply choose immediate availability—the probability of finding a required book available on the shelf. Variations include availability of the most popular books, satisfaction with the loan period and the waiting time, and some cost-based objectives. We believe that immediate availability is not an adequate measure of user satisfaction, which must include satisfaction with both the loan period and the maximum number of books that are allowed to be borrowed. Just as a user becomes dissatisfied when books are not available, he or she will be dissatisfied if the loan period is too short or the books he or she is allowed to borrow are too few. Thus, three measures of satisfaction are used in this article:

$s1$ = satisfaction with the loan period,
$s2$ = satisfaction with the maximum number of books allowed to be borrowed, and
$s3$ = satisfaction with book availability.

### 3.1. Model Assumptions

1. User satisfaction, for all three types, is defined to be a ratio of satisfied demands to total demands. This is consistent with approaches used in the literature. For example, Hindle and Buckland (1976) define satisfaction level as the *proportion* of demands which can be immediately satisfied.

2. User demands, for the number of books or the length of loan period, will not be affected by changes in the loan policy. If this is not the case, then according to Hindle and

Buckland (1976), one would expect the renewal probability to change with the loan period. It was mentioned earlier that, to the contrary, the length of the official loan period is seen to have negligible influence on the frequency of renewals.

3. As discussed previously, availability is affected by many factors, such as the loan period, duplication, and demand, which have been identified by Buckland (1975), Kantor (1976a), and Goehlert (1978). However, both Buckland and Goehlert conclude that the loan policy has, by far, the most significant impact on availability. Therefore, all other factors are assumed constant, and thus their effect is ignored, as we concentrate on the relationship between the loan policy and availability.

4. For the model to be realistic, $s3$ should be considered to have a more significant impact on overall user satisfaction than either $s1$ or $s2$. The level of dissatisfaction resulting from book unavailability should be greater than that resulting from not getting all the borrowing time or number of books desired. If books are not available, the user will not be concerned about how many to borrow or how long to keep them. Determining the relative weight of the three satisfaction components requires further research well beyond the scope of this article. Therefore, it will be assumed here that $s3$ is much more important than either $s1$ or $s2$.

### 3.2. Model Parameters

Let

- $I$ = upper limit on the official loan period in weeks
- $J$ = upper limit on the maximum number of books a user is allowed to borrow
- $i$ = length of the official (maximum) loan period in weeks, $i = 1, \ldots, I$
- $j$ = maximum number of books allowed to be borrowed by the same user, $j = 1, \ldots, J$
- $w$ = number of loan weeks desired by the user, $w = 1, \ldots, I$
- $b$ = number of books desired to be borrowed by the user, $b = 1, \ldots, J$
- $a$ = number of books available, out of those desired by the user, $a = 0, \ldots, b$.

Obviously, $s1$ (satisfaction with loan period $i$) is a function of $i$, and $s2$ (satisfaction with number of books $j$) is a function of $j$. Since availability is affected by the loan period and the number of books allowed, $s3$ (satisfaction with book availability) is a function of both $i$ and $j$. A user is partially satisfied if the loan policy does not allow him all the borrowing time or number of books he needs; in this case, we assume that satisfaction is the ratio of what is needed to what is allowed. On the other hand, a user is completely satisfied if the policy meets or exceeds his needs. Keeping in mind that satisfaction cannot exceed 100%, $s1$, $s2$, and $s3$ can be defined for each user as follows:

$$s1 = \min\left\{1, \frac{i}{w}\right\} \tag{1}$$

$$s2 = \min\left\{1, \frac{j}{b}\right\} \tag{2}$$

$$s3 = \frac{a}{b}. \tag{3}$$

Surveys of user needs must be conducted to determine user borrowing needs and current book availability. From these, probability distributions can be constructed for both the loan period and number of books required for borrowing. Given

$f_w$ = the probability that a user needs $w$ weeks
$g_b$ = the probability that a users needs $b$ books.

The average satisfaction with the loan period, $S1_i$, and with the number of books, $S2_j$, can be calculated for any given policy as follows:

$$S1_i = \sum_{w=1}^{I} f_w^* \min\left\{1, \frac{i}{w}\right\}, \; i = 1, \ldots, I \tag{4}$$

$$S2_j = \sum_{b=1}^{J} g_b^* \min\left\{1, \frac{j}{b}\right\}, j = 1, \ldots, J. \tag{5}$$

Another approach is used to calculate average availability $S3$. It is easier to calculate book unavailability ($S3' = 1 - S3$), since it is directly related to the loan period $i$ and the maximum number of books $j$. A logical assumption is that unavailability—the probability of not finding a needed book—depends on the average borrowing time $W_i$ times the average number of borrowed books $B_j$. If the needed time $w$ is less than the official loan period $i$, the user will keep books for $w$ weeks; if $w$ is greater than $i$, he/she will borrow them for only $i$ weeks. Similarly, if the number of needed books $b$ is less than the maximum number allowed $j$, the user will borrow $b$ books; if $b$ is greater than $j$, he will borrow only $j$ books. Noting that the loan period cannot exceed $i$, and that the number of books cannot exceed $j$, $W_i$ and $B_j$ can be estimated for each policy as shown below.

$$W_i = \sum_{w=1}^{I} f_w^* \min\{w, i\}, \; i = 1, \ldots, I \tag{6}$$

$$B_j = \sum_{b=1}^{J} g_b^* \min\{b, j\}, j = 1, \ldots, J \tag{7}$$

For each policy, average unavailability $S3'$ is assumed proportional to the product of the average borrowing time $W$ times the average number of borrowed books $B$. Let us define $\alpha$ as the average demand per book per week. If one book is checked out for one week, there will be an average

of $\alpha$ unavailabilities (unsatisfied demands). In general, if on average $B_j$ books are checked out for $W_i$ weeks, the average unavailability is $\alpha W_i B_j$. Thus, for each library, there is a demand constant $\alpha$ such that:

$$S3'_{ij} = \alpha W_i B_j,\ i = 1, \ldots, I,\ j = 1, \ldots, J. \quad (8)$$

### 3.3. Decision Variables

$$X_i = \begin{cases} 1 & \text{if the loan period is } i \text{ weeks} \\ 0 & \text{otherwise} \end{cases}, \quad i = 1, \ldots, I \quad (9)$$

$$Y_j = \begin{cases} 1 & \text{if the number of books is } j \text{ books} \\ 0 & \text{otherwise} \end{cases}, \quad j = 1, \ldots, J \quad (10)$$

### 3.4. Objective Functions

The objective is to maximize average total satisfaction, which is the aggregate of $S1$, $S2$, and $S3$. First, $S3$ is multiplied by the constant $C$ ($C > 1$) to indicate that it is much more important than $S1$ or $S3$. However, since $S3 = 1 - S3'$, maximizing the availability $S3$ is replaced by maximizing the negative of unavailability $S3'$.

$$\text{Maximize} \sum_{i=1}^{I} S1_i X_i + \sum_{j=1}^{J} S2_j Y_j - C \sum_{i=1}^{I} \sum_{j=1}^{J} S3'_{ij} X_i Y_j \quad (11)$$

or, from Equation 8:

$$\text{Maximize} \sum_{i=1}^{I} S1_i X_i + \sum_{j=1}^{J} S2_j Y_j - C\alpha \sum_{i=1}^{I} \sum_{j=1}^{J} W_i B_j X_i Y_j. \quad (12)$$

Here, $\alpha$ is a demand constant, defined by Equation 8, which must be calculated from the data for each library. First, calculate average unavailability $S3'$ for the current loan policy of $n$ weeks and $m$ books. For a sample of $K$ users, $S3'$ and $\alpha$ are calculated as:

$$S3'_{nm} = 1 - \left(\sum_{k=1}^{K} a_k / \sum_{k=1}^{K} b_k\right) \quad (13)$$

$$\alpha = \frac{S3'_{nm}}{W_n B_m}. \quad (14)$$

It can also be reasonably assumed, as indicated by Buckland (1972) and Shaw (1976b), that user needs for different loan periods have little effect on book return time, i.e., that users almost always wait until the books are due back. In this case, the average loan period $W_i$ is equal to the official loan period $i$, and the objective function becomes:

$$\text{Maximize} \sum_{i=1}^{I} S1_i X_i + \sum_{j=1}^{J} S2_j Y_j - C\alpha \sum_{i=1}^{I} \sum_{j=1}^{J} i B_j X_i Y_j. \quad (15)$$

From the data for the *current* loan policy of $n$ weeks and $m$ books, the constant $\alpha$ is now computed by:

$$\alpha = \frac{S3'_{nm}}{n B_m}. \quad (16)$$

Both objective functions 12 and 15 are nonlinear, since they involve the product of $X_i$ and $Y_j$. In order to solve by linear programming, the linearization technique developed by Watters (1967) for 0–1 variables must be applied first. Fortunately, this is not necessary, since the solution can be easily obtained by simple search.

### 3.5. Constraints

Since only one official loan period must be chosen, only one of the 0–1 $X_i$ variables must be equal to 1, while the others must be equal to 0.

$$\sum_{i=1}^{I} X_i = 1 \quad (17)$$

A similar constraint must be included to make only one choice for the maximum number of books.

$$\sum_{j=1}^{J} Y_j = 1 \quad (18)$$

## 4. Implementation Considerations

Unfortunately, computerized circulation data do not usually show user needs with respect to the loan policy, nor contain information on book availability. An exit poll must be conducted at the circulation checkout counter to obtain these data from users after they have been through the library; only then can they supply information regarding book availability. A convenient approach for doing this is by means of a short questionnaire that the user is requested to fill out at the time of checking out books. It is important to emphasize the fact to users that we want to know their

TABLE 1. Loan period demands and calculations.

| Maximum loan period in weeks $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| % Demand for $i$ weeks = $100 * f_i$ | 5 | 6 | 8 | 2 | 36 | 3 | 6 | 5 | 20 | 9 |
| % Satisfaction with $i$ weeks = $100 * S_i$ | 23.5 | 41.9 | 57.4 | 70.2 | 82.5 | 87.6 | 92.2 | 96.0 | 99.1 | 100 |
| Average borrowing time in weeks $W_i$ | 1 | 1.95 | 2.84 | 3.65 | 4.44 | 4.87 | 5.27 | 5.61 | 5.90 | 5.99 |

actual needs, regardless of the official loan policy currently in force.

The input parameters to the model must be obtained by surveying users of the library. For each user, data must be collected on the needed number of weeks $w$, needed number of books $b$, and number of needed books available $a$. In determining the frequency of user demands for given numbers of books $g_b$, the number of needed books $b$ must include books already on loan by the same user, reflecting total user needs. However, in calculating availability using Equations 3 or 13, books already on loan must be included, since we are interested only in immediate availability. The upper limits on the number of weeks $I$ and number of books $J$ must be set by the librarian, according to both user needs and practical considerations. Any given user needs that exceed any of these limits must be considered equal to the violated limit(s).

For obtaining the optimum solution, it is not necessary to use specialized integer programming software. Because of the simple 0–1 structure, it is more convenient to use spreadsheet packages such as Excel or Lotus. With simple spreadsheet techniques, one can develop an $I \times J$ matrix, in which each $(i, j)$ cell contains the following value for the objective function:

$$S1_i + S2_j - C\alpha W_i B_j. \tag{19}$$

The cell with the highest value is chosen. Let this be cell $(d, e)$, then the optimum loan policy calls for a maximum loan period of $d$ weeks and a maximum of $e$ books per user. Of course, $W_i$ must be set equal to $i$ for the second objective function 15. Although the integer programming model is not necessary for the solution, it provides a formal mathematical representation of the optimization problem. Moreover, it serves as a basis for more elaborate models with additional realistic considerations. For example, different loan policies for faculty, graduate students, and undergraduates can be optimized in an enlarged model.

## 5. Case Study

The proposed model was applied successfully at the central library of King Fahd University of Petroleum and Minerals in Saudi Arabia. There are three classes of users with three different loan policies: Faculty, graduate students, and undergraduate students. The study concentrated on the class of undergraduate students, as it was the largest group. The loan policy for undergraduates allows a maximum of seven books and 4 weeks. First, a questionnaire form was developed to obtain data on student needs and book availability. Students were asked to ignore the current loan regulations and express their actual needs by answering the following questions:

1. How many books do you *need* to borrow?
2. How many other books have you already borrowed and not yet returned?
3. Of the number needed (stated in Question 1), how many books were checked out by others?
4. How many weeks do you *need* to keep the borrowed books?

On the basis of students input and practical considerations, the maximum number to be considered of weeks $I$ and books $J$ were both set to 10. This agrees with Hindle and Buckland's (1976) survey which showed that loan periods in most university libraries do not exceed 10 weeks.

Table 1 shows the demand probabilities, percentage satisfaction, and average borrowing time for each official (maximum) loan period. Frequency of student needs for certain lengths of the borrowing time in weeks, taken from the survey, was used to obtain the demand probabilities. For each official loan period $i$, average percentage satisfaction with the number of weeks $S1_i$ was calculated using Equation 4. The average borrowing time in weeks $W_i$ was computed by Equation 6.

Table 2 shows the demand probabilities, percentage satisfaction, and average number of borrowed books, for each limit on the number of books. Frequency of student needs for certain numbers of books was used to obtain the demand probabilities. For each limit on the number of books $j$, average percentage satisfaction with the number books $S2_j$ was computed using Equation 5. The average number of borrowed books $B_j$ was calculated for each book limit $j$ by Equation 7.

### 5.1. Objective Function 12

In order to apply the model, constants must be calculated based on the present loan policy and current book availability. First, using Equation 13 on the survey data, current average unavailability $S3'$ was determined to be 22.5%,

TABLE 2. Number of books' demands and calculations.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Maximum no. of books allowed $j$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| % Demand for $j$ books = $100 * g_j$ | 6 | 6 | 15 | 17 | 25 | 7 | 13 | 3 | 2 | 6 |
| % Satisfaction with $j$ books = $100 * S2_j$ | 27.5 | 48.9 | 67.4 | 80.9 | 90.1 | 94.3 | 97.4 | 98.6 | 99.4 | 100 |
| Average no. of borrowed books $B_j$ | 1 | 1.94 | 2.82 | 3.55 | 4.11 | 4.42 | 4.66 | 4.77 | 4.85 | 4.91 |

which meant average availability is 77.5%. Tables 1 and 2 respectively show the average loan time and average number of borrowed books for the existing policy of 4 weeks and seven books ($W_4 = 3.65$, $B_7 = 4.66$). Equation 13 can now be used to calculate the demand constant for the first objective function 12.

$$\alpha = 0.225/(3.65) * (4.66) = 0.01323$$

Having calculated $\alpha$, we can determine average book availability $S3_{ij}$, the probability of finding a needed book, for each limit on the loan period $i$ and number of books $j$. First, average unavailability $S3'_{ij}$ is found by Equation 8, then $S3$ is set equal to $1 - S3'_{ij}$. The values obtained are displayed in Table 3.

Now, all the input parameters needed for the model have been calculated. Arbitrarily setting $C = 2$, the resulting binary integer programming model was then solved using the LINDO® linear programming package. The following optimal solution was obtained:

Maximum loan period $i$ = 7 weeks
Maximum number of books $j$ = 5 books
Percent average satisfaction = 100(S1+S2+S3)/3
= (92.2+90.1+72.3)/3
= 81.3%.

### 5.2. Objective Function 15

For the second objective function 15, the constant $\alpha$ is calculated by Equation 16, then average availability values $S3_{ij}$ are calculated as before and shown in Table 4.

$$\alpha = 0.225/4 * (4.66) = 0.01207$$

The optimal solution for the second objective, also obtained by LINDO®, is given below.

Maximum loan period $i$ = 7 weeks
Maximum number of books $j$ = 5 books
Percent average satisfaction = (92.2 + 90.1 + 71.9)/3
= 80.9%

### 5.3. Sensitivity Analysis

With both objective functions, the above solutions have been obtained by using the somewhat arbitrary value of (2.0) for $C$, the coefficient of $S3$, indicating that $S3$ is twice as significant as $S1$ or $S2$. For objective function 12, the above solution remains valid for values of this coefficient ranging from 2.0 to 2.2. For objective function 15, the solution is valid for coefficient values ranging from 1.8 to 2.1. This indicates that the solution is rather sensitive to the value of the coefficient. There is clearly a need to establish more reliable values for the coefficients of $S1$, $S2$, and $S3$. Therefore, a full-scale investigation of user opinions must be carried out to determine the relative influence of the three factors on overall satisfaction.

## 6. Comparison with Buckland's Values

It is interesting to note the remarkable similarity between the availability figures found by the new model, shown in Tables 3 and 4, and those determined by Buckland (1975).

TABLE 3. Percentage book availability $S3_{ij}$ for each policy with objective 12, assuming average borrowing time is $i$ weeks.

| | Books $j$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Weeks $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 98.7 | 97.4 | 96.3 | 95.3 | 94.6 | 94.2 | 93.8 | 93.7 | 93.6 | 93.5 |
| 2 | 97.4 | 95 | 92.7 | 90.8 | 89.4 | 88.6 | 88 | 87.7 | 87.5 | 87.3 |
| 3 | 96.2 | 92.7 | 89.4 | 86.7 | 84.6 | 83.4 | 82.5 | 82.1 | 81.8 | 81.6 |
| 4 | 95.2 | 90.6 | 86.4 | 82.9 | 80.2 | 78.7 | 77.5 | 77 | 76.6 | 76.3 |
| 5 | 94.1 | 88.6 | 83.4 | 79.2 | 75.9 | 74 | 72.3 | 72 | 71.5 | 71.2 |
| 6 | 93.6 | 87.5 | 81.8 | 77.1 | 73.5 | 71.5 | 70 | 69.3 | 68.8 | 68.4 |
| 7 | 93 | 86.5 | 80.3 | 75.3 | 71.4 | 69.2 | 67.5 | 66.8 | 66.2 | 65.8 |
| 8 | 92.6 | 85.6 | 79.1 | 73.7 | 69.5 | 67.2 | 65.4 | 64.6 | 64 | 63.6 |
| 9 | 92.2 | 84.9 | 78 | 72.3 | 67.9 | 65.5 | 63.6 | 62.8 | 62.2 | 61.7 |
| 10 | 92.1 | 84.6 | 77.7 | 71.9 | 67.4 | 65 | 63.1 | 62.2 | 61.6 | 61.1 |

TABLE 4. Percentage book availability $S3_{ij}$ for each policy with objective 15, assuming average borrowing time is $i$ weeks.

| | Books $j$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Weeks $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | 98.8 | 97.7 | 96.6 | 95.7 | 95 | 94.7 | 94.4 | 94.2 | 94.1 | 94.1 |
| 2 | 97.6 | 95.3 | 93.2 | 91.4 | 90.1 | 89.3 | 88.8 | 88.5 | 88.3 | 88.1 |
| 3 | 96.4 | 93 | 89.8 | 87.1 | 85.1 | 84 | 83.1 | 82.7 | 82.4 | 82.2 |
| 4 | 95.2 | 90.6 | 86.4 | 82.9 | 80.2 | 78.7 | 77.5 | 77 | 76.6 | 76.3 |
| 5 | 94 | 88.3 | 83 | 78.6 | 75.2 | 73.3 | 71.9 | 71.2 | 70.7 | 70.4 |
| 6 | 92.8 | 85.9 | 79.6 | 74.3 | 70.2 | 68 | 66.3 | 65.5 | 64.9 | 64.4 |
| 7 | 91.6 | 83.6 | 76.2 | 70 | 65.3 | 62.7 | 60.6 | 59.7 | 59 | 58.5 |
| 8 | 90.3 | 81.3 | 72.8 | 65.7 | 60.3 | 57.3 | 55 | 53.9 | 53.2 | 52.6 |
| 9 | 89.1 | 78.9 | 69.4 | 61.4 | 55.4 | 52 | 49.4 | 48.2 | 47.3 | 46.7 |
| 10 | 87.9 | 76.6 | 66 | 57.1 | 50.4 | 46.6 | 43.8 | 42.4 | 41.5 | 40.7 |

The availability values obtained by the two objective functions 12 and 15 were compared to those determined by Buckland for loan periods of 1, 2, 5, and 10 weeks. Table 5 shows the percentage availability values computed by the proposed model for the current policy of seven books, with the corresponding Buckland (1975, p. 92) values for popularity class $B$ one-copy loan policy.

It is not surprising to see that the second objective function yields the closest values to Buckland's. After all, the two models are based on the same assumption that users will always wait until the books are due back before returning them to the library. Naturally, complete agreement is not expected, since availability values calculated by the proposed model depend on the frequency distributions of user demands, and thus will vary from library to library. However, the similarity confirms that the assumptions made about availability are valid, and indicates that the model as a whole has a sound theoretical basis.

## 7. Conclusions

A new model for maximizing user satisfaction with a library's loan policy has been introduced. This integer programming model considers both elements of a loan policy: The loan period and the maximum number of books a user can borrow. User satisfaction is measured in terms of satisfaction with these two elements, as well as with book availability. By including these elements, the proposed model offers a more complete and realistic representation of the problem. The model has the advantage of easy solution, which can be obtained by simple spreadsheet software.

TABLE 5. Percentage availability values calculated by three methods.

| | Weeks | | | |
|---|---|---|---|---|
| Loan period | 1 | 2 | 5 | 10 |
| Objective 12: 7 books | 94 | 88 | 73 | 63 |
| Objective 15: 7 books | 94 | 89 | 72 | 44 |
| Buckland (1975, p. 92) | 94 | 86 | 62 | 44 |

A case study has been presented, illustrating how model parameters are calculated from user demand distributions. The values obtained compared well with previously published results. For future research, further investigation could be carried into the relative weights of the three measures of satisfaction mentioned above. Alternatively, loan policies for different groups, i.e., faculty, graduate and undergraduate students, could be simultaneously optimized. Furthermore, renewals and reservations (recalls), as well as duplication and multiple loan periods, could be included for a more comprehensive analysis.

## References


Bookstein, A. (1975). Optimal loan periods. *Information Processing and Management, 11,* 235–242.

Bruce, D. R. (1975). A Markov model to study the loan dynamics at a reserve-loan desk at a lending library. *Library Quarterly, 45,* 161–179.

Buckland, M. K. (1972). An operational research study of a variable loan and duplication policy at the University of Lancaster. *Library Quarterly, 42,* 97–106.

Buckland, M. K. (1975). *Book availability and the library user.* New York: Pergamon Press.

Burkhalter, B. R., & Race, P. A. (1968). Analysis of renewals and overdues and other factors influencing the optimal charge-out period. In B. R. Burkhalter (Ed.), *Case studies in systems analysis in a university library* (pp. 11–33). Metuchen, NJ: Scarecrow.

Burrell, Q. L. (1980). A simple stochastic model for library loans. *Journal of Documentation, 36,* 115–132.

Burrell, Q. L., & Fenton, M. R. (1994). A model for library book circulations incorporating loan period. *Journal of the American Society for Information Science, 45,* 101–116.

Goehlert, R. (1978). Book availability and delivery service. *The Journal of Academic Librarianship, 4,* 368–371.

Goyal, S. K. (1970). Applications of operational research to the problem of determining appropriate loan periods for periodicals. *Libri, 20*(1), 94–100.

Hindle, A., & Buckland, M. K. (1976). Toward an adaptive loan and duplication policy for a university library. In P. Brophy, M. K. Buckland, & A. Hindle (Eds.), *Reader in operations research for libraries* (pp. 69–76). Englewood, CO: Indian Head.

Hindle, A., & Worthington, D. (1980). Simple stochastic models for library loans. *Journal of Documentation, 36,* 209–213.

Kantor, P. B. (1976a). Availability analysis. *Journal of the American Society for Information Science, 27,* 311–319.

Kantor, P. B. (1976b). The library as an information utility in the university context: Evolution and measurement of service. *Journal of the American Society for Information Science, 27,* 100–112.

Kantor, P. B. (1979). A review of library operations research. *Library Research, 1,* 295–345.

Kraft, D. H., & Boyce, B. R. (1991). *Operations research for libraries and information agencies.* San Diego, CA: Academic Press.

Morse, P. M. (1968). *Library effectiveness: A systems approach.* Cambridge, MA: MIT Press.

Morse, P. M. (1979). A queueing theory Bayesian model for the circulation of books in a library. *Operations Research, 27,* 673–716.

Saracevic, T., Shaw, W. M., Jr., Kantor, P. B. (1977). Causes and dynamics of user frustration in an academic library. *College & Research Libraries, 38,* 7–18.

Shaw, W. M., Jr. (1976a). Library-user interface: A simulation of the circulation subsystem. *Information Processing & Management, 12,* 77–91.

Shaw, W. M., Jr. (1976b). Loan period distribution in academic libraries. *Information Processing & Management, 12,* 157–159.

Sinha, B. K., & Clelland, R. C. (1976). Modeling for the management of library collections. *Management Science, 22,* 547–557.

Watters, L. J. (1967). Reduction of integer polynomial programming problems to zero-one linear programming. *Operations Research, 15,* 1171–1174.

Yagello, V. E., & Guthrie, G. (1975). The effect of reduced loan periods on high use items. *College and Research Libraries, 36,* 411–414.

# On the Fusion of Documents from Multiple Collection Information Retrieval Systems

**Ronald R. Yager**
*Machine Intelligence Institute, Iona College, New Rochelle, NY 10801. E-mail: ryager@iona.edu*

**Alexander Rybalov**
*AT&T Bell Laboratories, 600 Mountain Ave., Murray Hill, NJ 07904*

**We investigate the problem of fusing collections of documents provided by multiple information retrieval systems. A parameterized approach is suggested in which a parameter determines how the documents in the individual collections are interleaved to form a fused list of documents. We then suggest a mechanism for learning this parameter.**

## 1. Introduction

An important problem is that of the retrieval of text documents, satisfying the requirements of a user. In Salton and McGill (1983), Salton (1989), Meadow (1992), and Frakes and Baeza-Yates (1992), a number of techniques to address this information retrieval problem are described. In some situations, especially with emergence of the Internet, users have at their disposal multiple independent sources in which to search for their documents. For example, in searching for some information about future stock prices, a user may search information bases of articles maintained by the *New York Times,* the *Wall Street Journal,* the *Chicago Tribune,* and other newspapers. Each one of these sources can provide a list of relevant articles which they published. Having these multiple sources raises the problem of fusing the results of searchers on these individual information bases (Bartell, Cottrell, & Belew, 1994; Belkin, Kantor, Cool, & Quatrain, 1993; Voorhees, Gupta, & Johnson-Laird, 1994, 1995). This problem is called the collection fusion problem. A particularly interesting solution to this problem has been suggested in Voorhees et al. (1994, 1995). One difficulty with the methodology used in Voorhees et al. (1994, 1995) is the nondeterminism of the fused list; the same query run twice in a row can result in a different ordering of the retrieved documents. This is due to the use of a probabilistic mechanism in the fusion process which leads to a randomness in the result. In this work, we look at the problem of fusing ordered lists of documents and suggest a number of alternative approaches, all of which are deterministic.

## 2. Collection Fusion Problem

In Voorhees et al. (1994, 1995), the collection fusion problem is formally described. Assume that we have a group of information servers. Here we have in mind, as a server, a collection of text documents, like that which could be maintained by the *New York Times* consisting of all the articles that have appeared in it. We shall denote these as $S_i$ for $i = 1$ to $m$, $m$ being the number of servers. We assume each server contains a unique collection of documents, a document appears in only one server. Furthermore, it is assumed that each server has its own distinct searching mechanism. Using these searching mechanisms, for a given query $Q$, each server can associate a score with each document contained in it and thus provide an ordered list of relevant documents in response to the query. The problem we are interested in solving is one in which we desire to obtain an ordered list of the $N$ most relevant documents, where these documents can come from any of the servers. The difficulty in accomplishing this task resides in the fact that the scoring mechanisms used by the individual servers may be incomparable, and hence we cannot just simply take the $N$ highest scoring documents across all servers. An approach for solving this problem has been suggested (Voorhees et al., 1994, 1995) which can essentially be seen to consist of a three-step process:

1. Pass the query $Q$ to each of the individual servers and obtain a score for each of the documents. This scoring allows us to obtain, from each of the servers, a list of documents ordered by their relevancy to the query.
2. Based upon the nature of the query $Q$, determine how many documents should be contributed by the server $S_i$.

Received April 7, 1997; revised August 19, 1997 and September 29, 1997; accepted October 31, 1997.

We shall denote this as $N_i$; it is assumed that $\Sigma_{i=1}^{m} N_i = N$. Combining this with step one, we obtain, from each server, an ordered list, $L_i$, consisting of its $N_i$ highest scoring documents.

3. Fuse the individual lists, $L_1, L_2, \ldots L_m$, to obtain an ordered list of the $N$ desired documents which we denote as $L$.

It should be clear that the issues relating specifically to the collection fusion problem arise with respect to steps two and three, the determination of the number of elements to select from each collection and the fusion of the lists; step one can be implemented by any of a number of approaches that are now available (Frakes & Baeza, 1992; Meadow, 1992; Salton, 1989; Salton & McGill, 1983).

Step two requires some comparison between the individual information servers regarding their potential for containing documents relevant to the question being asked. The solution to this problem is highly information intense and can be seen to require empirical studies on the contents of the different information bases, with respect to the query being posed. Formally, in order to determine how many documents we should take from each server under the query $Q$, given the constraint that we should take a total of $N$ documents, all we need is the relative proportion of the documents which we should take from source $S_i$ with respect to another server, say $S_1$. That is, we need to obtain for $i = 2$ to $m$, a value $\beta_i$, based upon the query, where $\beta_i$ is the proportion of documents to be contributed by $S_i$, compared to that of $S_1$. Once having these $\beta_i$, we know that

$$N_i = \beta_i N_1, \text{ for } i = 2 \text{ to } m$$

and since

$$\sum_{i=1}^{m} N_i = N$$

we get $N = N_1 + \beta_2 N_1 + \beta_2 N_1 + \cdots \beta_m N_1 = N_1(1 + \Sigma_{i=2}^{m} \beta_i)$. This can be solved for $N_1 = N(1 + \Sigma_{i=2}^{m} \beta_i)^{-1}$ which gives us all the $N_i$.

In this work we shall concentrate on the solution to the problem in the third step, fusing the individual collection to get one collection, for as we shall see in the following, the solution suggested in Voorhees et al. (1994, 1995) has problems.

## 3. Ordered List Fusion

We now focus on the process involved in step three of the preceding algorithm; we have a collection of $m$ ordered lists, $L_i$, of distinct elements, and we desire to fuse these lists into one ordered list $L$. We should note that this problem has some connection with the classic work done by Arrow (1951) on social choice.

In forming the combined list $L$ from the individual lists, we require that one property be satisfied, *intra-list consistency.* This condition requires that if $x$ and $y$ are two documents appearing in list $L_i$ and if $x$ is preferred (is higher up on $L_i$) over $y$ in this list, i.e., $x >_i y$, then it is required that in the fused list $L$ $x$ is higher than $y$, i.e., $x >_L y$.

In order to combine these individual ordered lists, we need some guiding imperative. However, on the surface, there appears no natural principle for guiding us in the task of combining these lists. Essentially, there appears only one distinction between the lists: The number of elements in each list.

An approach to the fusion of these lists has been suggested and tested (Voorhees et al., 1994, 1995) and which, according to the studies, worked well. The method uses a probabilistic mechanism to determine each of the elements to be placed on $L$. In the method suggested, the list $L$ is constructed in descending order. In particular, using a random experiment, one selects one of the contributing lists, and then selects the top available element from that list and places it in the next position in $L$. One continues in this manner until all contributing list are depleted. The basic mechanism for selecting the next element to be added to $L$ is as follows.

Assume after selecting the first $r$ documents in $L$, one has $n_i$ documents left in each of the individual lists to be fused. One then associates with each list $L_i$, a probability

$$p_i = \frac{n_i}{\Sigma_{j=1}^{m} n_j}.$$

One then associates with each collection $L_i$, an interval

$$I_i = [a_i, b_i)$$

where

$$a_1 = 0$$

$$a_i = b_{i-1} \quad i = 2, \cdots m$$

$$b_i = a_i + p_i \quad i = 1, \cdots m.$$

One next generates a random number $ran \in [0, 1]$. If $ran \in I_i$, one selects the top element in list $L_i$ to be added to $L$. This process can be seen as a biased random generation. It is clear that this method satisfies the property of intra-list consistency. In Voorhees et al. (1994, 1995), it is shown empirically that this procedure results in good ordering. However, this approach has one fundamental flaw. Because of the probabilistic mechanism used, each time we implement this procedure, we could get a different ordering in $L$, i.e., *it is non-deterministic.* In the environment of information retrieval, this is unacceptable. For if a user starts going down the list to check the documents, stops, and then later

resubmits the query, the user will get a different list. This, of course, causes confusion and anxiety.

The spirit for fusing the individual lists used in Voorhees et al. (1994, 1995) can be seen to be based upon the following imperative:

*The more elements left in an individual list, <u>the more likely it is to contain</u> the best remaining element.*

We should note that the spirit of the above approach seems reasonable. First, it is efficient in that it uses as the basic discriminating factor the only distinguishing feature between the contributing lists. the number of elements currently in each of the lists. Furthermore, since, as indicated in step two, the original number of elements in each list is determined by some measure of the overall relevancy of the documents in that server to the query, the usage of the number of elements in the individual lists to reflect its potential for having the next best document is consistent. In addition, the empirical results described in Voorhees et al. (1994, 1995) seem to validate the use of the number of elements as a good source of discriminating information. However, the non-determinism is unacceptable in real systems.

In the following, we suggest an alternative approach, which is essentially in the same spirit, but leads to a deterministic ordering. Assume, after selecting the first $r$ documents in $L$, we have $n_i$ elements left in list $L_i$. We now associate with each of the individual collections, a value

$$V_i = \frac{n_i}{\sum_{j=1}^{m} n_j}.$$

Note that $V_i = p_i$. We then select, as the next element to add to $L$, the top element left in the collection with the highest value for $V_i$. We also note that this approach uses the same information to determine which contributing list is the best source for the next document, without using a non-deterministic, random mechanism.

One problem must be addressed, situations in which two or more lists are tied with respect to their $V$ value. In order to address this problem, we shall use the following deterministic tie-breaking procedure based upon an indexing of the information sources. We index the individual collections in descending order on the basis of the total number of documents they are going to contribute to the fused collection; thus, if $S_i$ and $S_j$ are two servers, we shall assume $i < j$ if $N_i > N_j$. Furthermore, if there are any ties with respect to the number of documents that individual collections are going to contribute, we shall adjudicate these by alphabetically ordering the tied collections; thus if $S_i$ and $S_j$ are such that $N_i = N_j$, we shall assume $i < j$ if the name of $S_i$ appears higher in the alphabet then $S_j$. Using this indexing, if at some point in the fusion process there is a tie with respect to the $V$ values, we select the next document from the collection with the smallest index. Using this tie-breaking procedure results in our fusion process always being deterministic for a given query.

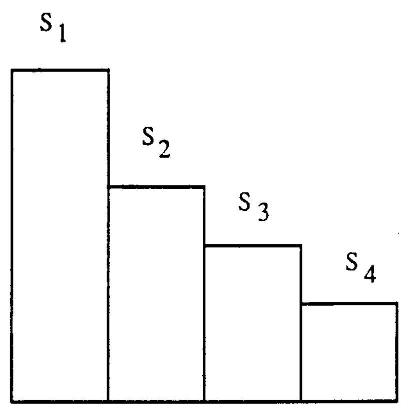


FIG. 1. Collection to be fused.

Figure 1 helps provide an understanding of the performance of this fusion approach.

We take the first $N_1$–$N_2$ elements from collection one. We then take, from collections one and two, the top elements in round-robin fashion,[1] until each of these collections have been reduced to containing $N_3$ elements. We next take, from collections 1, 2, and 3, the top elements in round-robin fashion until the number of elements in each of these collections have been reduced to containing $N_4$ elements. Finally, we take documents from collections 1, 2, 3, and 4, in round-robin fashion, until all collections have been emptied.

This approach can be seen to be one in which the potential of list $L_i$ providing the next element for the fused list is determined by the number of documents remaining in $L_i$. Essentially, the preceding imperative becomes a deterministic one, that is:

*The more elements left in an individual list, <u>then it contains</u> the best element.*

Here we have replaced *likely* with a kind of certainty that allows us to avoid the non-determinism of the random approach.

In the scheme just introduced, we use, as the value to determine from which collection to take the next document, a function proportional to the number of documents left in a collection. Another approach is to consider the number of elements we have taken from the collection. In particular, here we consider an alternative imperative which says that *the fewer elements taken from a collection, the more likely it is to contain the best element.* In order to accomplish this, we can associate with each of the individual collections a value

$$V_i = \frac{\text{Min}(1, (N_i - g_i))}{g_i}$$

[1] By round-robin fashion, we mean that we select the top element from $S_1$, then the top element from $S_2$.

where $N_i$ is the number of elements originally in collection $S_i$, and $g_i$ is the number of elements taken from collection $S_i$. We place the top element from the collection with the highest score on the next position on the overall ranking. If there are ties with respect to this score, we use our previously described tie-breaking procedure, taking from the one collection among the tied collections that has the lowest index value. Using $\wedge$ for Min and $\vee$ for Max, we can express this as

$$V_i = \frac{1 \wedge ((N_i - g_i) \vee 0)}{g_i}.$$

We note that if a collection has no element left, $N_i - g_i = 0$, then $V_i = 0$. On the other hand, when $N_i - g_i > 0$, then we see that $V_i = 1/g_i$. In the case when all the collections have at least one element left, we can express this selection function in a normalized form

$$U_i = \frac{\frac{1}{g_i}}{\sum_{j=1}^{m} \frac{1}{g_j}},$$

where $U_i \in [0, 1]$. If we let $\Pi_{k \neq j}\, g_k$ equal the product of all the $g$s except $g_j$, then we can express the above as

$$U_i = \frac{\Pi_{k \neq i}\, g_k}{\sum_{j=1}^{m} (\Pi_{k \neq j}\, g_k}.$$

It should be noted that the approach described here is very much in the same spirit as a random selection in which each collection is assumed to have an equally likely chance of having the best document available. We can see this as follows. Initially $g_i = 0$ for all $i$, hence $U_i$ must be the same for all $i$, $U_i = 1/m$. Using our tie-breaking mechanism, we select the top element for each of the tied collections in round-robin fashion based on their index. After this selection, we now have $g_i = 1$ for all $i$, and again $U_i = 1/m$. Again, we use our tie-breaking mechanism. It should be apparent that using this approach means that the $U_i$ are always equal for non-empty collections.

Figure 2 helps to provide an understanding of the workings of this approach.

In this approach, we essentially push all of the individual orderings up so that their top elements are equal. We then will take the top elements from collections 1, 2, 3, and 4 in round-robin fashion. We continue until all the elements in $S_4$ are exhausted. We then take the top elements in 1, 2, and 3 in round-robin fashion until all the elements in $S_3$ are exhausted. We continue in this manner until all elements are exhausted.

The underlying assumption with this approach can be viewed as one in which the top element in each of the

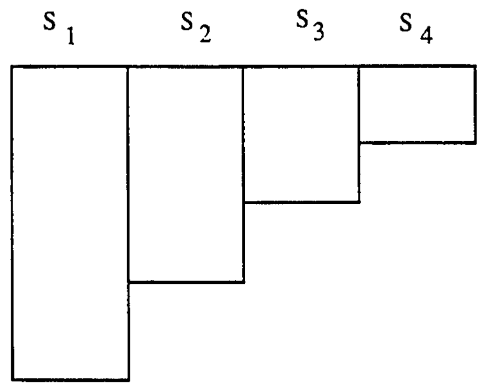


FIG. 2. View of second fusion procedure.

collections are assumed equal regarding the possibility of being the next best available document.

An alternative manifestation of this fusion agenda can be had using the function

$$V_i = -g_i,$$

where again $g_i$ is the number of elements taken from collection. Here again, we take from the collection having the highest $V_i$ value, unless it is empty.

Thus far, we have considered two alternative deterministic approaches to the fusion of the individual lists. The first approach takes the next element from the top of the list having the most elements remaining. This approach essentially makes some distinction between the lists based upon the number of elements remaining. The second approach, while formally selecting from the list which has contributed the least number of elements to $L$, essentially makes no distinction between the lists and says the top elements in each list are just as good. It is a deterministic manifestation of randomly selecting from each list.

As can be best seen from the preceding Figures 1 and 2, there exists a view of the two methodologies for list fusion based on the relative arrangement of the individual lists. In the two approaches discussed so far, one consisted of a moving down of collections, so that the worst elements are on the same level (see Fig. 1), while the other can be seen as one in which all collections are equalized by moving the smaller ones up, so that all top elements are equal (see Fig. 2). In Figure 3, we show an alternative way of fusing the individual collection documents.

In this approach, we essentially center the collections and fuse them. In order to formally implement this imperative, we associate, with each of the collections to be fused, a value

$$V_i = \tfrac{1}{2} N_i - g_i$$

where again, $N_i$ is the number of documents initially in collection $i$, and $g_i$ is the number of documents already removed from collection $i$. Using this function, we again

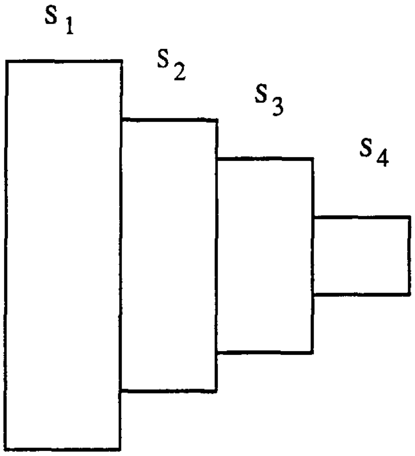


FIG. 3. Centralizing fusion approach.

take the top element from the collection having the largest $V$ value, if that collection is not already depleted, and place it in the next highest position in our fused collection. If two or more collections are tied with respect to their $V_i$ value, we use our usual tie-breaking procedure. The following example illustrates the performance of this methodology.

*Example*

Assume we have four collections having respectively 9, 5, 3, and 1 elements to be fused. The lists of elements from each of these collections are given below:

$$\begin{array}{lllllllll} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 & a_9 \\ b_1 & b_2 & b_3 & b_4 & b_5 \\ c_1 & c_2 & c_3 \\ d_1. \end{array}$$

Initially, the scores are: $V_1 = 4.5$, $V_2 = 2.5$, $V_3 = 1.5$, and $V_4 = 5$. We then remove $a_1$ from collection one and place it on top of the fused list. This gives us new scores of $V_1 = 3.5$, $V_2 = 2.5$, $V_3 = 1.5$, and $V_4 = 0.5$. Again, we remove the top element from collection one, $a_2$, and place it in the next top position in the fused list. This gives us new scores of $V_1 = 2.5$, $V_2 = 2.5$, $V_3 = 1.5$, and $V_4 = 0.5$. We now take the top elements from collections 1 and 2, $a_3$ and $b_1$, and place them in the fused collection. We note we put $a_3$ first because of our tie-breaking rule. This gives us new scores: $V_1 = 1.5$, $V_2 = 1.5$, $V_3 = 1.5$, and $V_4 = 0.5$. From this, we place elements $a_4$, $b_2$, and $c_1$ in the fused collection. Following this procedure, we get the following fused list:

$$a_1a_2a_3b_1a_4b_2c_1a_5b_3c_2d_1a_6b_4c_3a_7b_5a_8a_9.$$

We see that the fused list essentially is a centralized type of fusion. Immediately below, we illustrate the same fusion but show tied elements in the same column, which makes more apparent the centralizing aspect of this fusion mechanism:

$$\begin{array}{ccccccccc} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 & a_9 \\ & & b_1 & b_2 & b_3 & b_4 & b_5 & & \\ & & & c_1 & c_2 & c_3 & & & \\ & & & & d_1. & & & & \end{array}$$

It is interesting to look at a generalization of this centralizing fusion approach. Consider the function

$$V_i = \alpha N_i - g_i$$

where $\alpha \in [0, 1]$. If we let $\alpha = 1$, we get

$$V_i = N_i - g_i = n_i$$

where $n_i$ is the number of documents left in collection $S_i$. This is essentially the same evaluation function used in the first fusion method we introduced. If we let $\alpha = 0$, we get as our evaluation function $V_i = -g_i$. This is essentially the same valuation function used in the second fusion method we introduced. If we let $\alpha = 0.5$, we get this last centralizing fusion method. Thus, we see that by selecting $\alpha$, we are essentially changing the relationship of shorter lists with respect to the longer ones. It is interesting to note that since $N_i = n_i + g_i$, we can express this general formulation as

$$V_i = \alpha(n_i + g_i) - g_i$$

$$V_i = \alpha n_i - (1 - \alpha)g_i \quad \alpha \in [0, 1].$$

This general fusion function can be seen as a weighted average of the number of elements remaining in each stack and the negative of the number of elements taken from each collection.

In the following, we illustrate the fusion of the four collections in the previous example using $\alpha = 0$, $\alpha = 0.25$, $\alpha = 0.5$, $\alpha = 0.75$, and $\alpha = 1$. (Tied elements are shown in the same column)

1. $\alpha=0$

$$\begin{array}{lllllllll} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 & a_9 \\ b_1 & b_2 & b_3 & b_4 & b_5 \\ c_1 & c_2 & c_3 \\ d_1 \end{array}$$

2. $\alpha = 0.25$

$$\begin{array}{cccccccccccc} a_1 & a_2 & c_1 & a_3 & c_2 & a_4 & c_3 & a_5 & a_6 & a_7 & a_8 & a_9 \\ & b_1 & & b_2 & & b_3 & & b_4 & b_5 & & & \\ & & & d_1 & & & & & & & & \end{array}$$

3. $\alpha = .5$

$$\begin{array}{ccccccccc} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 & a_9 \\ & & b_1 & b_2 & b_3 & b_4 & b_5 & & \\ & & & c_1 & c_2 & c_3 & & & \\ & & & & d_1 & & & & \end{array}$$

4. $\alpha = .75$

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $c_1$ | $a_6$ | $c_2$ | $a_7$ | $c_3$ | $a_8$ | $a_9$ |
| | | | $b_1$ | $b_2$ | | $b_3$ | | $b_4$ | | $b_5$ | |
| | | | | | | | | $d_1$ | | | |

5. $\alpha = 1$

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $a_7$ | $a_8$ |
| | | | $b_1$ | $b_2$ | $b_3$ | $b_4$ | $b_5$ |
| | | | | | $c_1$ | $c_2$ | $c_3$ |
| | | | | | | | $d_1$ |

From these examples, we clearly see the shifting of the resulting fused ordering as we change $\alpha$. Essentially, we see that the larger the value of $\alpha$, the more we delay using the shorter stacks, while for smaller $\alpha$'s, we hasten the use of the shorter stacks.

One semantic that we can associate with the parameter $\alpha$ is that of being a measure of the degree to which one believes the collection with the most remaining elements contains the best of the unfused documents. On the other hand $(1 - \alpha)$ can be interpreted as the degree to which one believes that any collection can contain the best element.

It should be noted that $V_i = \alpha N_i - g_i$ has two interesting properties. The first is that $V_i$ only depends on attributes associated with the collection $S_i$; it is independent of the attributes associated with any of the other lists. Second, $V_i$ is monotonically decreasing in $g_i$. These properties have important implications in tied environments. Assume, at some point in the process, the biggest scores are $V_i = V_j = V_k$ where $i < j < k$. According to our tie-breaking mechanism, we select the top element in $L_i$. This selection, because of the monotonicity property noted above, reduces $V_i$, however, because of the independence property, this selection leaves all other values the same. Thus, we now have as our maximum scores $V_j = V_k$. We would then select the element from $L_j$. Finally, we select the top element in $L_k$. The implication here is that, at any point in the process, if we have a tie for the biggest score, we can select the top element in *all* the tied collections at once and place them in the fused list, based upon their index.

Actually, we can provide a unique value associated with each of the documents to be fused under this agenda. Let $x_{ij}$ be the $j$th document in the $i$th collection. Again, $N_i$ is the total number of objects to be contributed by the $i$th collection. We can associate with each document, $x_{ij}$, a value $V_{ij}$ and place the documents in the fused list based upon this value, noting that the higher the value, the higher the position in the fused list. We recall that for the top element in collection $S_i$, its score is $V_i = \alpha N_i - g_i$, where $g_i$ is the number of elements already removed from this collection. Furthermore, we note that if $x_{ij}$ is the $j$th element in a collection, when it comes to the top, $j - 1$ elements will have already been removed. Thus the score associated with this element, $V_{ij}$, will always be

$$V_{ij} = \alpha N_i - (j - 1) = \alpha N_i + 1 - j.$$

We should note that documents from the same list will always have distinct values, $V_{ij} \neq V_{ij}$ for $j$ Of course, it is possible to have ties between elements from different lists. In this case, we use our tie-breaking procedure to adjudicate these ties. Using this approach, we can easily form the overall fused list from the individual lists in a deterministic way.

## 4. Learning the Fusion Parameter

The parameterization of our selection function with the introduction of $\alpha$ gives us considerable freedom. It allows us to decide upon an appropriate choice for the value of $\alpha$. One way to obtain $\alpha$ is to use empirical observations to learn the value of $\alpha$. In the following, we shall describe an approach to the learning of $\alpha$ based upon observations.

An observation consists of a query, $Q$, which generates a set of ordered lists, $L_1, L_2, \cdots L_m$, one from each of the information sources. In addition, we shall assume that there exists a final fused list $\hat{L}$. This final fused list is provided by a user who integrates the individual lists based on their perception of the satisfaction of each of the documents to the query.[2] We shall assume that we have a collection of $p$ observations. We shall indicate each observation by $\mathfrak{D}_i$, thus

$$\mathfrak{D}_i = [Q, L_1, \ldots L_m, \hat{L}].$$

Note that $Q$, $L_1, \cdots L_m$, and $\hat{L}$ are unique to each observation.

In order to learn the value of $\alpha$, we now consider a smaller component which we call an *action* and which we define as follows. Assume that we have an observation ˆi. $\hat{L}$ is constructed by the user taking an available element from one of the individual lists and placing it in the next available position in $\hat{L}$. We shall call this process an action. The construction of $\hat{L}$ is made up of a collection of actions, if there are $N$ elements in $\hat{L}$, then we have $N$ actions. Thus, at each step in the construction of $\hat{L}$ from the individual lists, the user implements an action.

The following example illustrates this idea.

*Example*

| $L_1$ | $L_2$ | $L_3$ | $\hat{L}$ |
|---|---|---|---|
| $a_1$ | $b_1$ | $c_1$ | $a_1$ |
| $a_2$ | $b_2$ | $c_2$ | $b_1$ |
| $a_3$ | $b_3$ | | $a_2$ |
| $a_4$ | | | $a_3$ |
| $a_5$ | | | $c_1$ |
| | | | $b_2$ |
| | | | $b_3$ |
| | | | $a_4$ |
| | | | $a_5$ |
| | | | $c_2$ |

[2] It is possible that this list $\hat{L}$ may not satisfy the intra-list consistency property. However, we shall see our approach is indifferent to this possibility.

In this observation, action 1 is to place $a_1$ in $\hat{L}$, action 2 is to place $b_1$ in $\hat{L}$, and action 3 is to place $a_2$ in $\hat{L}$.

As we indicated earlier, one semantic that can be associated with $\alpha$ is to interpret it as the degree to which an action is to select an element from the collection with the most remaining elements. Thus, for the purpose of learning $\alpha$, the aspect of an action of interest is whether the element was selected from one of the collections having the most remaining elements. In particular, what we need to obtain from each observation is the count of the number of actions in which we have individual collections having different numbers of elements, and the count of the number of times in these situations we select the document from one of the lists with the most elements. We do not consider situations in which all the individual lists have the same number of available elements. The following example, using the data of the proceeding example, illustrates this process.

*Example*

| Step | $n_1$ | $n_2$ | $n_3$ | Action |
|---|---|---|---|---|
| 1 | 5 | 3 | 2 | $L_1$ |
| 2 | 4 | 3 | 2 | $L_2$ |
| 3 | 4 | 2 | 2 | $L_1$ |
| 4 | 3 | 2 | 2 | $L_1$ |
| 5 | 2 | 2 | 2 | $L_3$ |
| 6 | 2 | 2 | 1 | $L_2$ |
| 7 | 2 | 1 | 1 | $L_2$ |
| 8 | 2 | 0 | 1 | $L_1$ |
| 9 | 1 | 0 | 1 | $L_1$ |
| 10 | 1 | 0 | 0 | $L_1$ |

We see that in steps 1, 2, 3, 4, 6, 7, and 8, we had a solution in which one of the non-depleted collections, had less elements than the other ones, thus the total number of relevant actions (RAs) is #RA = 7. We also see that in actions 1, 3, 4, 6, and 8, the expert selected their document from one of those lists having the maximal available elements, thus the number of affirmative actions (#AA) is 5.

We now are in a position to provide a method for obtaining the measure $\alpha$ from these observations. Assume we have $p$ observations, let #RA($i$) indicate the number of relevant actions in the $i$th observation, and #AA($i$) be the number of affirmative actions. Then, an estimate for $\alpha$ is

$$\alpha = \frac{\sum_{i=1}^{p} \#\mathrm{AA}(i)}{\sum_{i=1}^{p} \#\mathrm{RA}(i)}.$$

## 5. Proportional Approach to List Fusion

We will now introduce a different agenda for fusing the individual collections based upon the proportion of elements.[3] The basic idea of this approach can be seen in an example in which we have four collections containing 8, 4, 2, and 1 documents, respectively. Under this approach, we would take two from the first collection, then one from the second, and again two from the first than one from the second and one from the third, etc. In the following, we provide the final list:

$$\begin{array}{cccccccc}
a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 \\
 & b_1 & & b_2 & & b_3 & & b_4 \\
 & & & c_1 & & & & c_2 \\
 & & & & & & & d_1.
\end{array}$$

We now introduce a formal mechanism which implements this fusion mechanism. We associate with each collection a value

$$V_i = \frac{n_i - 1}{N_i}$$

where $N_i$ is the number of documents originally in that collection, and $n_i$ is the number of documents remaining in that collection at the current step in the fusion process. We then select the top element from the non-empty collection having the largest $V_i$ value. Ties are adjudicated by the usual tie-breaking method described earlier. Because we are always taking the top element from each collection, we can, in this case, also assign a unique value to each of the documents to be fused. If $x_{ij}$ is the $j$th document in the $i$th collection, the unique value associated with this document is

$$V_{ij} = \frac{N_i - (j-1) - 1}{N_i} = \frac{N_i - j}{N_i}.$$

To illustrate this approach, we consider the following two examples. In each case, the number in parenthesis is the value of $V_{ij}$.

*Example*

| $L_1$ | $L_2$ | $L_3$ | $L_4$ |
|---|---|---|---|
| $a_1$ (7/8) | $b_1$ (3/4) | $c_1$ (1/2) | $d_1$ (0) |
| $a_2$ (6/8) | $b_2$ (2/4) | $c_2$ (0) | |
| $a_3$ (5/8) | $b_3$ (1/4) | | |
| $a_4$ (4/8) | $b_4$ (0) | | |
| $a_5$ (3/8) | | | |
| $a_6$ (2/8) | | | |
| $a_7$ (1/8) | | | |
| $a_8$ (0/8) | | | |

From this, we get the following fused list:

$$\begin{array}{cccccccc}
a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 \\
 & b_2 & & b_2 & & b_3 & & b_4 \\
 & & & c_1 & & & & c_2 \\
 & & & & & & & d_1.
\end{array}$$

A second example, using this fusion method, is the case of four lists having 10, 6, 4, and 2 documents, respectively.

[3] We acknowledge our debt to Henrik L. Larsen, of Roskilde University in Denmark, for suggesting this type of fusion.

*Example*

| $L_1$ | $L_2$ | $L_3$ | $L_4$ |
|---|---|---|---|
| $a_1$ (.9) | $b_1$ (.83) | $c_1$ (.75) | $d_1$ (0) |
| $a_2$ (.8) | $b_2$ (.67) | $c_2$ (.5) | $d_2$ (0) |
| $a_3$ (.7) | $b_3$ (.5) | $c_3$ (.25) | |
| $a_4$ (.6) | $b_4$ (.33) | $c_4$ (0) | |
| $a_5$ (.5) | $b_5$ (.16) | | |
| $a_6$ (.4) | $b_6$ (0) | | |
| $a_7$ (.3) | | | |
| $a_8$ (.2) | | | |
| $a_9$ (.1) | | | |
| $a_{10}$ 0 | | | |

In this case, the fused list is:

$$\begin{array}{l}a_1b_1a_2c_1a_3b_2a_4\ a_5a_6b_4a_7c_3a_8b_5a_9\ a_{10}\\ \qquad\qquad\qquad\qquad\qquad\qquad b_3 \qquad\quad b_6\\ \qquad\qquad\qquad\qquad\qquad\qquad c_2 \qquad\quad c_4\\ \qquad\qquad\qquad\qquad\qquad\qquad d_1 \qquad\quad d_2.\end{array}$$

## 6. Conclusion

We have investigated the problem of fusing collections of documents provided by multiple information servers. A parameterized approach has been suggested in which the parameter determines how the documents in the individual collections are interleaved. Associating this parameter with the action that it is best to select a document from the collection with the most elements, we introduced a mechanism for learning the parameter.

## 7. References


Arrow, K. J. (1951). *Social choice and individual values.* New York: Wiley.

Bartell, B. T., Cottrell, G. W., & Belew, R. K. (1994). Automatic combination of multiple ranked retrieval systems. *Proceedings of the 17th International Conference on Research and Development in Information Retrieval.*

Belkin, N. J., Kantor, P., Cool, C., & Quatrain, R. (1993). Query combination and data fusion for information retrieval. *Proceedings of the Second Text Retrieval Conference (TREC-2)* (pp. 35–44).

Frakes, W. B., & Baeza-Yates, R. (1992). *Information retrieval: Data structures and algorithms.* Upper Saddle River, NJ: Prentice-Hall.

Meadow, C. T. (1992). *Text information retrieval systems.* New York: Academic Press.

Salton, G. (1989). *Automatic text processing.* Reading, MA: Addison-Wesley.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval.* New York: McGraw-Hill.

Voorhees, E. M., Gupta, N. K., & Johnson-Laird, B. (1994). The collection fusion problem. *Proceedings of the Third Text Retrieval (TREC-3) Conference* (pp. 95–104).

Voorhees, E. M., Gupta, N. K., & Johnson-Laird, B. (1995). Learning collection fusion strategies. *Proceedings of the 18th ACM SIGIR Conference on Research and Development in Information Retrieval,* Seattle, (pp. 172–179).

# Indexing and Access for Digital Libraries and the Internet: Human, Database, and Domain Factors

**Marcia J. Bates**
*Department of Information Studies, University of California, Los Angeles, Los Angeles, California 90095-1520*
*E-mail: mjbates@ucla.edu*

**Discussion in the research community and among the general public regarding content indexing (especially subject indexing) and access to digital resources, especially on the Internet, has underutilized research on a variety of factors that are important in the design of such access mechanisms. Some of these factors and issues are reviewed and implications drawn for information system design in the era of electronic access. Specifically the following are discussed: *Human factors:* Subject searching vs. indexing, multiple terms of access, folk classification, basic-level terms, and folk access; *Database factors:* Bradford's Law, vocabulary scalability, the Resnikoff-Dolby 30:1 Rule; *Domain factors:* Role of domain in indexing.**

## Introduction

### *Objectives*

In the current era of digital resources and the Internet, system design for information retrieval has burst onto the stage of the public consciousness in a way never seen before. Almost overnight, people who had never thought about information retrieval are now musing on how to get the best results from their query on a World Wide Web search engine, or from a remote library catalog or digital library resource. At the same time, and under the same stimulus, experts in a variety of fields cognate to information science—such as cognitive science, computational liguistics, and artificial intelligence—are taking up information retrieval questions from their perspectives.

In the meantime, those of us in information science, where information retrieval from recorded sources (as distinct from mental information retrieval) has long been a core, if not *the* core, concern, are faced with a unique mix of challenges. Information science has long been a field understaffed with researchers. Where some fields have 10 researchers working on any given question, we have often had one researcher for 10 questions. A promising research result from an information science study may languish for years before a second person takes up the question and builds on the earlier work. (This is not universal in the field; some questions are well studied.) As a consequence of this understaffing, we know a lot about many elements of information retrieval, but often in a fragmented and underdeveloped way.

At the same time, what we do know is of great value. Years of experience, not only with research but also with application in dozens of information-industry companies, have given information scientists a deep understanding about information retrieval that is missing in the larger world of science and information technology.

So at this particular historical juncture in the development of information retrieval research and practice, I believe there are a number of both research results and experience-based bits of knowledge that information scientists need to be reminded of, and non-information scientists need to be informed of—information scientists because the fragmentation and understaffing in our field has made it difficult to see and build on all the relevant elements at any one time, and people outside of information science because these results are unknown to them, or at least unknown to them in the information retrieval implications.

My purpose here is to draw attention to that learning and those research results associated with indexing and access to information, which have been relatively under-utilized by those both inside and outside of information science. These are results that seem to me to have great potential and/or importance in information system design, and for further research.

Making such a selection is, of course, a matter of judgment. I believe that the material below offers the possibility of enriching design and, when studied further, enriching our

Received January 31, 1997; revised August 28, 1997; accepted November 7, 1997.

understanding of human interactions with information in automated environments.

### *Fully Automated Subject Access*

Before beginning to review that knowledge, we must first address a prior issue. Many people in our society, including many in the Internet and digital resources environments, assume that subject access to digital resources is a problem that has been solved, or is about to be solved, with a few more small modifications of current full-text indexing systems. Therefore, why worry about the various factors addressed in this article?

There are at least two reasons. First, the human, domain, and other factors would still operate in a fully automated environment, and need to be dealt with to optimize the effectiveness of information retrieval (IR) systems. Whatever information systems we develop, human beings still will come in the same basic model; products of human activity, such as databases, still will have the same statistical properties, and so on. As should become evident, failure to work with these factors will almost certainly diminish the resulting product.

Second, it may take longer than we think to develop fully automated systems. At conferences, researchers present their new system and say, "We are 70% there with our prototype system. Just a little more work, and we will have it solved." This happens because, indeed, it is not difficult to get that first 70% in retrieval systems—especially with small prototype systems.

The last 30%, however, is infinitely more difficult. Researchers have been making the latter discovery as well for at least the last 30 years. (Many of the retrieval formulas that have recently been tried were first used in the 1950s and 1960s. See, e.g., Luhn, 1957; Stevens, 1965.) Information retrieval has looked deceptively simple to generations of newcomers to the field. But IR involves language and cognitive processing, and is therefore as difficult to automate as language translation and other language processes based on real-world knowledge, which researchers have been trying to automate virtually since the invention of the computer. Information retrieval also poses serious scalability problems; small prototype systems are often *not* like their larger cousins. Further, user needs vary not just from one time to another, but from one subject domain to another. Optimal indexing and retrieval mechanisms may vary substantially from field to field.

We can do an enormous number of powerful things with computers, but effective, completely automated indexing and access to textual and text-linked databases eludes us still, just as 100% perfect automatic translation does, among other things. Meanwhile, a lot of other IR capabilities go undeveloped, in deference to the collective assumption that we will soon find a way to make it possible for computers to do everything needed in information retrieval.

The human side of the IR process needs attention too. The really sophisticated use of computers will require designs shaped much more in relation to how human minds and information needs actually function, not to how formal, analytical models might assume they do. Attention to these points will be productive for effective IR, no matter how soon, or whether, we find a way to develop good, fully automated, IR.

## Human Factors

### *Subject Searching vs. Indexing*

It is commonly assumed that indexing and searching are mirror images of each other. In indexing, the contents are described or represented, or, in full-text searching, indicative words or phrases are matched or otherwise identified. On the searching side, the user formulates a statement of the query. Then these two representations, of document and of query, are matched to retrieve the results.

But, in fact, this is only superficially a symmetrical relationship. *The user's experience is phenomenologically different from the indexer's experience.* The user's task is to describe something that, by definition, he or she does not know (cf. Belkin, 1982). (Knowledge specifically of what is wanted would lead to a "known-item" search.) The user, in effect, describes the fringes of a gap in knowledge, and can only guess what the "filler" for the gap would look like. Or, the user describes a broader, more general *topic area* than the specific question of interest, and says, in effect, "Get me some stuff that falls in this general area and I'll pick what looks good to me." Usually, the user has no tools available to help with that problem of describing the fringes of the gap, or the broader subject area.

In many cases, the problem for the user is even more difficult than indicated above. In years of studies, Kuhlthau (1993) has documented that the very process of coming to know what one wants in the first place is seldom straightforward. In a search of any complexity, as for a student term paper, one does not so much "pick a topic," as is usually assumed, but rather discovers, develops, and shapes it over time through exploration in materials in an area of interest. One may use an information system at any point in this gradually developing process. The need may evolve and shift at each stage of developing knowledge of the subject. (See also my own model of the evolving and changing information need—Bates, 1989a.) Use of an information system early in a project will naturally come out of a much less well-specified and articulated information need—yet the searcher must nonetheless find a way to get the information system to respond helpfully.

The indexer, on the other hand, has the record in hand. It is all there in front of him or her. There is no gap. Here, ideally, the challenge for the indexer is to try to anticipate what terms people with information gaps of various descriptions might search for in those cases where the record in hand would, in fact, go part way in satisfying the user's information need. This can be seen to be a very peculiar challenge, when one thinks about it. What kinds of information needs would people have that might lead them to want some information that this record would, in fact, provide?

As Harter (1992) points out, discovering that a particular article is relevant to one's concerns, and therefore a good find, does not necessarily mean that the article is "about" the same subject as one's originating interest. As he notes, regarding his own article on the concept of psychological relevance: "The present article [on psychological relevance] may be found relevant, by some readers, to the topics of designing and evaluating information retrieval systems, and to bibliometrics; I hope that it will. However, this article is not *about* these topics" (p. 603, emphasis in the original). Conceivably, infinitely many queries could be satisfied by the record in hand. Imagining even the more likely ones is a major challenge. (See extensive discussions in Ellis, 1996; Green & Bean, 1995; O'Connor, 1996; Soergel, 1985; and Wilson, 1968.)

But in fact, historically, and often still today, catalogers and indexers do not index on the basis of these infinitely many possible anticipated needs. Instead, and perhaps much more practically, they simply *index what is in the record.* (See also discussion in Fidel, 1994.) In other words, they attempt to provide the most careful and accurate possible description or representation of the contents of the record. This situation differs in many ways from the phenomenological circumstances of the user. We should not be surprised, then, if the user and the indexer use different terminology to describe the record, or, more generally, conceptualize the nature and character of the record differently.

For the indexer, there is no mystery. The record is known, visible before him or her. Factual information can be checked, directly and immediately, to create an absolutely accurate record. The user, on the other hand, is seeking something unknown, about which only guesses can be made. What happens in retrieval if the searcher's guesses have little (or big) inaccuracies in them, and do not match the precise and accurate description provided by the indexer?

Further, the indexer is experienced with the indexing system and vocabulary. Over the years, with any given system, fine distinctions are worked out regarding when one term is to be used and when another for two closely related concepts. Indexers create rules to cover these debatable situations. Eventually, mastery of these rules of application comes to constitute a substantial body of expertise in itself. For example, the subject cataloging manual for the *Library of Congress Subject Headings* (Library of Congress, 1996) runs to two volumes and hundreds of pages (Library of Congress, 1991– ). *This manual is not the same as the subject heading listings.* It does not consist so much of general indexing principles as it does of rules for applying individual headings or types of headings. For example, under the subject "Strikes and Lockouts," section H2100 of the manual, four pages of instructions are given for creating headings for strikes in individual industries and for individually-named strikes of various types.

Thesaural indexing systems also frequently have "scope notes" that tell the indexer how to decide which term to use under debatable circumstances.

Often, these rules are not available to searchers, but even when they are, the naive searcher—and that includes Ph.D.s, as long as they are naive about indexing—will usually not realize there are any ambiguities or problems with a term, and will not feel a need to check it. The user has in mind the sense of a term that interests him or her, not the other senses that the indexer is aware of. Only upon retrieving false drops will the user realize there is even any problem.

In short, the user almost always knows less about the indexing issues in a topic area than the indexer does. The user approaches the system with an information need that may be formulated out the of the first words that come to mind. (See Markey, 1984b, on the many queries she found that "could be categorized as 'whatever popped into the searcher's mind'" p. 70.) Consequently, the user's input is liable not to be a good match with the indexer's labeling, which is derived from years of experience and analysis. The indexer, on the other hand, cannot undo his/her far greater knowledge of the indexing issues. After indexers have been at work for one day, or attended one training session, they know more, and have thought more, about indexing issues than even the most highly educated typical user has. Already, an expertise gap is forming between the user and the indexer that nearly guarantees some mismatches between user search terms and indexing terms on records.

The same phenomenological gap will hold true for the match between the system user and bodies of data that have been automatically indexed by some algorithm as well. The creator of the algorithm has likewise thought more and experimented more with the retrieval effects of various algorithms than the typical user has, before selecting the particular algorithm made available in a given system. Yet, at the same time, the full statistical consequences of such algorithms can never be anticipated for every circumstance.

Results achieved through such algorithms will seldom dovetail exactly with what humans would do in similar circumstances. So, the result is a peculiar mix of expert human understanding of indexing with non-sentient statistical techniques, to produce a result that is never just like interacting with another human would be. Again, we have a

phenomenologically different experience for the researcher/designer on one side, and the user on the other. Further, no retrieval algorithm has been found to be anywhere near perfectly suitable (more on this later), yet the principles by which the algorithm is designed are seldom made fully available to end users—just as the indexing principles seldom are—so the user could try to find ways around inadequacies.

Still another factor is likely to operate in the behavior of catalogers, indexers, and system designers for manual and automated systems. The information professional has an understandable desire to create, in an indexing vocabulary, classification system, or statistical algorithm, a beautiful edifice. He or she wants a system that is consistent in its internal structure, that is logical and rigorous, that can be defended among other professionals, as well as meet all the usual expectations for creations of human endeavor. After years of working with problems of description, the creators of such systems have become aware of every problem area in the system, and have determinedly found some solution for each such problem. Therefore, the better developed the typical system, the more arcane its fine distinctions and rules are likely to be, and the less likely to match the unconsidered, inchoate attempts of the average user to find material of interest.

This is by no means to suggest that IR systems should be inchoate or unconsidered! Instead, the question is a different one that is usually assumed. That question should not be: "How can we produce the most elegant, rigorous, complete system of indexing or classification?," but rather, "How can we produce a system whose front-end feels natural to and compatible with the searcher, and which, by whatever infinitely clever internal means we devise, helps the searcher find his or her way to the desired information?"

Such clever internal means may include having distinct means of access and indexing. The two functions of indexing and access are not one and the same. It is not necessary to require the user to input the "right" indexing term, which the system in turn uses to search directly on record indexing. The design of the access component of a system can be different from the design of the indexing component, provided these two are appropriately linked.

Thus, the user, with only vague, poorly thought through ideas initially of what is wanted, can be helped, with the right access mechanism, to find his or her way into the portions of the database that have been carefully and rigorously organized. It is unreasonable to demand of system users that they develop the expert indexer's expertise. At the same time, it is unreasonable to demand of the indexers that they be sloppy and fail to note critical, and useful, distinctions. By developing an access mechanism—a user-friendly front-end—for the user, and linking it to the indexing, indexers can maintain standards, and users can benefit from the intellectual work that has gone into the system of description, without having to be experts themselves. (For further discussion, see Bates, 1986a, 1990.)

### *Multiple Terms of Access*

On this next matter there is a huge body of available research. Several people have written extensively about it (Bates, 1986a, 1989b; Furnas, Landauer, Dumais, & Gomez, 1983; Gomez, Lochbaum, & Landauer, 1990), and yet the results from these studies are so (apparently) counterintuitive that little has been done to act on this information in IR system design. It is as if, collectively, we just cannot believe, and, therefore do not act upon, this data. A few examples of these results will be described here; otherwise the reader is encouraged to explore the full range of available data on this matter in the above-cited references.

*In study after study, across a wide range of environments, it has been found that for any target topic, people will use a very wide range of different terms, and no one of those terms will occur very frequently.* These variants can be morphological (forest, forests), syntactic (forest management, management of forests) and semantic (forest, woods).

One example result can be found in Saracevic and Kantor (1988). In a carefully designed and controlled study on real queries being searched online by experienced searchers, when search formulations by pairs of searchers *for the identical query* were compared, in only 1.5% of the 800 comparisons were the search formulations identical. In 56% of the comparisons, the overlap in terms used was 25% or less; in 94% of the comparisons, the overlap was 60% or less (p. 204).

In another study, by Lilley (1954), in which 340 library students were given books and asked to suggest subject headings for them, they produced an average of 62 different headings for each of the six test books. Most of Lilley's examples were simple, the easiest being *The Complete Dog Book,* for which the correct heading was "Dogs." By my calculation, the most frequent term suggested by Lilley's students averaged 29% of total mentions across the six books.

Dozens of indexer consistency studies have also shown that even trained experts in indexing still produce a surprisingly wide array of terms within the context of indexing rules in a subject description system (Leonard, 1977; Markey, 1984a). See references for many other examples of this pattern in many different environments.

To check out this pattern on the World Wide Web, a couple of small trial samples were run. The first topic, searched with the Infoseek search engine, is one of the best known topics in the social sciences, one that is generally described by an established set of terms: *The effects of television violence on children.* This query was searched on five different expressions that varied only minimally. The search was not varied at all by use of different search

capabilities in Infoseek, such as quotation marks, brackets, or hyphens to change the searching algorithm. Only the words themselves were altered—slightly. (Change in word order alone did not alter retrievals.) These were the first five searches run:

violent TV children
children television violence
media violence children
violent media children
children TV violence

As is standard, each search produced 10 addresses and associated descriptions as the first response for the query, for a total of 50 responses. Each of these queries could easily have been input by a person interested in the identical topic. If each query yielded the same results, there would be only 10 different entries—the same 10—across all five searches. Yet comparison of these hits found 23 different entries among the 50, and in varying order on the screen. For instance, the response labeled "Teen Violence: The Myths and the Realities," appeared as #1 for the query "violent media children" and #10 for "violent TV children." The search was then extended a little farther afield to a query on "mass media effects children." That query yielded nine new sites among the 10 retrieved, for a total of 32 different sites—instead of the 10 that might have been predicted—across the six searches.

The previous example varied the search words in small, almost trivial ways. The variation could easily have been much greater, while still reflecting the same interests from the searcher. Let us suppose, for example, that someone interested in freedom of speech issues on the Internet enters this query:

+"freedom of speech" +Internet

This time the search was run on the Alta Vista search engine. The plus signs signal the system that the terms so marked must be present in the retrieved record and the quotation marks require the contained words to be found as a phrase, rather than as individual, possibly separated, words in the record. So, once again, we have a query that is about as straightforward as possible; both terms, as written above, must be present.

Let us suppose, however, that three other people, interested in the very same topic, happen to think of it in just a little different way, and, using the identical system search capabilities, so that only the vocabulary differs, they enter, respectively:

+"First Amendment" +Web
+"free speech" +cyberspace
+"intellectual freedom" +Net

All four of these queries were run, in rapid succession, on Alta Vista. The first screen of 10 retrievals in each case was compared to the first 10 for the other three queries. The number of different addresses could vary from 10 (same set across all four queries) to 40 (completely different set of 10 retrievals for each of the four queries). Result: There were 40 different addresses altogether for the four queries. Not a single entry in any of the retrieved sets appeared in any of the other sets.

Next, the search was expanded by combining the eight different terms in all the logical combinations, that is, each first term combined with each second term (free speech and Internet, free speech and Web, First Amendment and cyberspace, etc.). There are 16 such orders, for a total of 160 "slots" for addresses on first 10 retrievals in each case. All 16 of these combinations could easily have been input by a person interested in the very same issue (as well as dozens, if not hundreds, of other combinations and small variations on the component terms).

The result: Out of the 160 slots, 138 unique different entries were produced. Thus, if each of the 16 queries had been entered by a different person, each person would have missed 128 other "top 10" entries on essentially the same topic, not to mention the additional results that could be produced by the dozens of other terminological and search syntax variations possible on this topic. In sum, the data from these small tests of the World Wide Web conform well with all the other data we have about the wide range of vocabulary people use to describe information and to search on information.

If 85 or 90% of users employed the same term for a given topic, and only the remainder used an idiosyncratic variety of other terms, we could, with a moderate amount of comfort, endeavor to satisfy just the 85 or 90%, by finding that most popular term and using it in indexing the topic. But description of information by people just does not work this way. Even the most frequently used term for a topic is employed by a minority of people. There are generally a large number of terms used, many with non-trivial numbers of uses, and yet no one term is used by most searchers or indexers.

For search engines designed as *browsers* this may be a good thing. The slight variations yield different results sets for searchers, and thus spread around the hits better across the possible sites, thereby promoting serendipity. But for people making a directed search, it is illusory to think that entering that single just-right formulation of the query, if one can only find it, will retrieve the best sites, nicely ranked, with the best matches first.

Under these circumstances, any simple assumption about one-to-one matching between query and database terms does not hold. In million-item databases, even spelling errors will usually retrieve something, and reasonable, correctly spelled terms will often retrieve a great many hits. (In one of my search engine queries, I accidentally input the misspelling "chidlren" instead of "children." The first four retrievals all had "chidlren" in the title.) Users may simply not realize that the 300 hits they get on a search—far more

than they really want anyway—are actually a small minority of the 10,000 records available on their topic, some of which may be far more useful to the user than any of the 300 actually retrieved.

Another possible reason why we have not readily absorbed this counter-intuitive study data: Interaction with a system is often compared to a conversation. Whether or not a person consciously thinks of the interaction that way, the unconscious assumptions can be presumed to derive from our mental model of conversations, because that is, quite simply, the principal kind of interaction model we language-using humans come equipped with. In a conversation, if I say "forest" and you say "forests," or I say "forest management" and you say "management of forests," or I say "forest" and you say "woods," we do not normally even notice that different terms are used between us. We both understand what the other says, each member of the example pairs of terms taps into the same area of understanding in our minds, and we proceed quite happily and satisfactorily with our conversation. It does not occur to us that we routinely use this variety and are still understood. Computer matching algorithms, of course, usually do not generally build in this variety, except for some stemming, because it does not occur to us that we need it.

Experienced online database searchers have long understood the need for variety in vocabulary when they do a thorough search. In the early days of online searching, searchers would carefully identify descriptors from the thesaurus of the database they were searching. The better designed the thesaurus and indexing system, the more useful this practice is. However, searchers soon realized that, in many cases where high recall was wanted, the best retrieval set would come from using as many different terms and term variants as possible, including the official descriptors. They would do this by scanning several thesauri from the subject area of the database and entering all the relevant terms they could find, whether or not they were official descriptors in the target database.

In some cases, where they had frequent need to search a certain topic, or a concept element within a topic, they would develop a "hedge," a sometimes-lengthy list of OR'd terms, which they would store and call up from the database vendor as needed. (See, e.g., Klatt, 1994.)

Sara Knapp, one of the pioneers in the online searching area, has published an unusual thesaurus—not the kind used by indexers to identify the best term to index with, but one that searchers can use to cover the many terms needed for a thorough search in an area (Knapp, 1993). Figure 1 displays the same topic, "Child development," as it appears in a conventional thesaurus, the *Thesaurus of ERIC Descriptors* (Houston, 1995), and in Knapp's searcher thesaurus. It can be seen that Knapp's thesaurus provides far more variants on a core concept than the conventional indexer thesaurus does, including likely different term endings and possible good Boolean combinations.

A popular approach in IR research has been to develop ranking algorithms, so that the user is not swamped with hundreds of undifferentiated hits in response to a query. Ranking will help with the 300 items that are retrieved on the term-that-came-to-mind for the user—but what about the dozens of other terms and term variations that would also retrieve useful material (some of it far better) for the searcher as well? IR system design must take into account these well-attested characteristics of human search term use and matching, or continue to create systems that operate in ignorance of how human linguistic interaction in searching actually functions.

## All Subject Vocabulary Terms Are Not Equal

There is considerable suggestive evidence that the human mind processes certain classes of terms differently from others, i.e., that certain terms are privileged in comparison with others. If so, we can ask how we might take advantage of these characteristics in IR system design.

### *Folk classification*

There is a substantial body of linguistic and anthropological research into what are called "folk classifications," the sets of categories used by various cultures for plants, animals, colors, etc. (See, e.g., Brown, 1984; Ellen & Reason, 1979; Raven, Berlin, & Breedlove, 1971.) As Raven, Berlin, and Breedlove note: "In all languages, recognition is given to naturally occurring groupings of organisms. These groupings appear to be treated as psychologically discontinuous units in nature and are easily recognizable" (p. 1210). These they refer to as *taxa.* Across many cultures, these groupings have been shown to fall into a few class types, which the authors label, going from general to specific: "Unique beginner, life form, generic, specific, varietal" (p. 1210).

Of these, the "generic" is pivotally important, containing readily recognizable forms such as "monkey" or "bear." The more specific forms, "specific" and "varietal," are generally few in number, and "can be recognized linguistically in that they are commonly labeled in a binomial or trinomial format that includes the name of the generic or specific to which they belong" (p. 1210), such as—in Western cultural terms—"howler monkey" or "Alaskan grizzly bear."

The generics are so readily recognized because there are many physical discontinuities that distinguish one from another (Brown, 1984, p. 11). For instance, the shape and appearance of a cow (generic level) is very different from the shape and appearance of a fox (generic level), while the same for a Guernsey cow (specific or varietal) is not so different from the shape and appearance of a Hereford cow (specific or varietal).

**CHILD DEVELOPMENT** *Jul. 1966*
CIJE: 4101 RIE: 3775 GC: 120
BT Individual Development
RT Child Behavior
Child Development Centers
Child Development Specialists
Child Health
Child Language
Child Rearing
Child Responsibility
Children
Delayed Speech
Developmental Delays
Developmental Stages
Developmental Tasks
Failure to Thrive
Family Environment
Parenthood Education
Piagetian Theory

**Child development. Child development.** *Choose from:* **child(ren,hood), juvenile, pediatric, infant(s), bab(y,ies), neonat(e,es, al), adolescen(t,ce), psychosocial, psychosexual, emotional, personality, physical, cognitive, moral, intellectual, social, language, autonomy, motor, sexual, psychophysiological, psychomotor, speech** *with:* **development(al), develop(ed,ing), growth, matur(ed,ing,ation), immatur(e, ity), transition(s,al), stage(s).** *Consider also:* **oral, anal, phallic, genital** *with:* **stage(s).** *See also* **Adolescent development; Age differences; Autism; Bonding (emotional); Child development disorders; Child language; Child nutrition; Childrearing; Cognitive development; Delayed development; Developmental disabilities; Developmental stages; Early childhood development; Egocentrism; Emotional development; Failure to thrive (psychosocial); Home environment; Human development; Individual differences; Infant development; Life cycle; Motor development; Object relations; Oedipus complex; Perceptual development; Physical development; Psychomotor development; Socialization.**

FIG. 1. Comparison of "child development" in two thesauri. *Note.* Entry on left reprinted from *Thesaurus of ERIC Descriptors, 13th Edition,* p. 43, edited by James E. Houston © 1995 by The Oryx Press. Used with permission from The Oryx Press, 4041 N. Central Ave., Suite 700, Phoenix, AZ, 85012. 800-279-6799. http://www.oryxpress.com. *Note.* Entry on right reprinted from *The Contemporary Thesaurus of Social Science Terms and Synonyms: A Guide for Natural Language Computer Searching,* p. 52, compiled and edited by Sara D. Knapp © 1993 by above-named publisher, and reprinted with the publisher's permission.

Raven, Berlin, and Breedlove (1971) go on to say:

> Folk taxonomies all over the world are shallow hierarchically and comprise a strictly limited number of generic taxa ranging from about 250 to 800 forms applied to plants and a similar number applied to animals. These numbers are consistent, regardless of the richness of the environment in which the particular people live. (p. 1213)

So the English-speaking Californian will have a set of names for plants or animals that falls in the same number range as the Yapese speaker on an island in Micronesia or an Eskimo in Canada. (And, no, in contrast to what is commonly stated, even in serious scientific works, Eskimos do *not* have hundreds of words for "snow," trouncing English with just one or two words for the same phenomenon. Read Pinker, 1994, p. 64ff, as he skewers what he calls the "Great Eskimo Vocabulary Hoax.")

A striking discovery in this research on folk classifications has been the observation that formal scientific taxonomies of biological taxa started out, and remained for some time, folk classifications in structure. Several generations of biological taxonomies, including, above all, that of Linnaeus, resemble folk classifications on several structural features. "In broad outlines . . . the system of Linnaeus was a codification of the folk taxonomy of a particular area of Europe." (Raven, Berlin, & Breedlove, 1971, p. 1211).

Both linguistic and anthropological research are more and more frequently demonstrating common underlying patterns across cultures and languages, whatever the particulars of the expression of those patterns in given cultures. I believe, even, that an argument can be made that the Dewey Decimal Classification also betrays some of these same folk classification characteristics.

But whether or not Dewey or other formally constituted classifications retain folk classification roots, we can certainly expect the average information system user to approach that system with a mind that creates—and expects to find—classifications with such characteristics.

What would these classifications look like? Research was not found that links this research in anthropology with IR research, but we might at least expect, and test for, the pattern described above: Shallow hierarchy (few levels), with the generic level of prime significance. In any one area that generic level might contain 250 to 800 terms. If we were to find this pattern in information-seeking behavior—that is, people are found to favor use of systems with categories numbered in this range—then the implications for design of access in automated information systems are clear.

*Basic level terms*

Other research has been conducted in psychology that also supports the "not all vocabulary terms are equal" idea. Eleanor Rosch and her colleagues (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Rosch, 1978) have done extensive research supporting the idea that, in natural language vocabulary grouping and use, there are what she calls "basic level" terms. As a psychologist, Rosch has endeavored to identify patterns, not to compare across cultures, but rather to discover as fundamental human patterns of cognitive processing. She, too, identified the importance of natural discontinuities. She states:

> A working assumption of the research on basic objects is that (1) in the perceived world, information-rich bundles of perceptual and functional attributes occur that form natural discontinuities, and that (2) basic cuts in categorization are made at these discontinuities. (Rosch, 1978, p. 31)

In a long series of studies, she found that these basic level terms are the ones that are learned first by children, and are easiest to process and identify by adults (Rosch et al., 1976; Rosch, 1978). Both Rosch (1978, p. 32) and Berlin (1978, p. 24) note the likely correspondence between the generic level in folk classifications and the "basic level" in the psychological studies.

She (Rosch et al., 1976) and Newport and Bellugi (1978) even found that the centrality of basic level terms appears in sign language for the deaf. Research throughout the Rosch studies was done on basic level terms (e.g., chair) plus superordinate (e.g., furniture) and subordinate (e.g., kitchen chair) terms. Newport and Bellugi (1978) state, regarding the signs in American Sign Language, that "superordinate and subordinate signs are usually derived from signs at the basic level: They contain basic-level signs as their components. In short, Rosch's basic level is formally basic in American Sign Language" (p. 52).

When we turn to information science, we do not know if there are basic level search terms used by information system users. Rosch's work has tantalized the field for many years, but it is not easy to find a way to identify basic level terms when people are talking about topics of interest in information seeking in contrast to Rosch's very carefully controlled laboratory studies. (See Brown's (1995) efforts to study subject terms from a more psychological perspective.) However, the task should not be insurmountable. Looking at how college students search for material on a topic when assigned to write a paper might reveal patterns in the level of generality of their search terms, for instance. Much more research is needed in this area.

However, it would seem to be a reasonable assumption that we would find that certain terms come more commonly to mind for the general user of the Web browser or database, terms that are neither the broadest nor narrowest (i.e., that are "basic level"), and if we could find a way to identify and use those terms for access, system users would make a match much more readily in their searches.

*Folk access*

The discussion above was about "folk classification." We in information science might also talk about "folk access." There is some evidence that there are identifiable patterns in what people expect of information systems, though much more needs to be studied on this matter. In their study of online catalog design, Kaske and Sanders (1980a, 1980b) found that focus groups responded positively to the idea of some sort of classification tree being available to users at the beginning of their search. My dissertation research (Bates, 1977, p. 371) also uncovered a substantial amount of spontaneous classification behavior on the part of the student participants in a study of the *Library of Congress Subject Headings* (Library of Congress, 1996). That is, a substantial minority of the students frequently used broad terms with narrower topical subdivisions, despite the fact that such topical subdivisions were forbidden at that time in indexing with Library of Congress subject headings.

There are many problems with attempting to offer classified access in the design of subject access systems in manual and automated information systems, and I am not suggesting that access systems should necessarily offer that approach, at least in the conventional sense of classified access in library systems. Rather, I suggest that some sort of design that enables people to approach a digital information system with the *feel* of classified access may be helpful for users. There will be more on this matter later in the article.

Finally, there is another body of evidence about the nature of "folk access." It is an anecdotal truism among reference librarians that users frequently do not say what they really want when they come to a reference desk, but rather ask broader questions. The user who really wants to find out Jimmy Hoffa's birth date instead asks where the books on labor unions are.

In an empirical study of reference interviews, Lynch (1978) found that 13% of the several hundred interviews she observed involved shifts from an initial presenting question to a different actual question. Although Lynch did not analyze these shifts in terms of breadth, she did find that the great majority of the 13% were cases in which the user presented either a directional question ("Where are the . . . ") or a holdings question ("Do you have [a specific item]"), and it turned out that the person really wanted information on a subject not specific to a particular book.

Thomas Eichman has argued persuasively (1978) that this is a quite reasonable approach for users to use in the actual circumstances of the reference interview, and should be expected rather than bemoaned by reference librarians. Usually, the person with a query is approaching a total

stranger, and has no idea how much the librarian might know about the subject of interest, or how helpful she will be able or willing to be. Initial interactions constitute "phatic" communication, wherein the user is opening up a channel of communication, sizing up and getting to know the person on the other side of the desk.

Eichman argues that in asking an initial broad or general question, the user is much more likely to be able to establish some point of contact with the stranger behind the desk. Consider: If the user asks, "Where are the psychology books?" or "Do you have so-and-so's *Introduction to Psychology?*", can he not have a higher expectation of the librarian's being able to answer than if he asks for his true information need, a book that can tell him about the Purkinje Effect?

Let us suppose, further, that the librarian has, indeed, never heard of the Purkinje Effect. If the user asks for this point-blank, the librarian cannot immediately help him, without asking more questions. The user anticipates that there will be a socially awkward moment, as he gets a blank look from the librarian. If, on the other hand, he starts with the question about psychology books, and the conversation proceeds well, he and the librarian will move through a conversation that makes it clear that the user is interested in visual perception, and the Purkinje Effect is a topic within that area. The librarian, having been given this necessary context, can now, in fact, direct him to the right books in the psychology section, even though she still may not know specifically what the Purkinje Effect is.

Starting general, then, would seem to be an effective strategy in the information-seeking conversation between librarian and user. Librarians sometimes fail to recognize the value of this approach from the user, because they (the librarians) are focusing on the technical, professional matter of getting the question answered, just as some physicians bypass the phatic and contextual interactions in a patient interview to get at what they see as the "real business" of diagnosis and treatment. Unfortunately, the patient does not yield up private matters to total strangers so easily, and the professional's "short" route to the technical content of the interview often founders on its failure to recognize how the interview feels to the user/patient.

In broad outline, this movement from general to specific in the course of the reference interaction sounds rather like the pattern of broad term/narrower term noted above in some of the student subject headings. It is as though they expect to connect with the information system by starting broad and working down to the specific: Psychology—Visual Perception—Purkinje Effect. If, as assumed, people carry conversation patterns over into interaction with information systems, then enabling the human user to survey the territory, to size up the system's offerings before settling down directly to business might equip the searcher with a better understanding of the system's capabilities, and therefore enable that searcher to submit queries better matched to the strengths of the system.

Another part of my dissertation analyzed the match between the student-generated terms and the library subject headings actually applied to the test books. The students without library training more frequently used terms *broader* than the actual heading applied than they used terms narrower than the correct heading (Bates, 1977, p. 370ff).

Other data in my study supported the idea that use of broader terms was a wise (if probably unconscious) strategy on the part of the students. One might expect, a priori, that broader search terms would be easier to match to index terms than narrower ones, because they are usually shorter, simpler, and more generally agreed-upon in phrasing. This indeed proved to be the case. That half of the students who erred most by using broader terms were compared with the half who used the fewest broad terms. The match rates of the former were higher (Bates, 1977, p. 373).

In a study of different design, Brooks (1995) also confirmed my result, finding that "narrower descriptors are harder for subjects to match to bibliographic records than the other two types" (p. 107). The "other two types" were what he called "topical-level" and "broader." The former were terms assigned to a record by an indexer; broader and narrower terms were those that were broader and narrower in relation to the assigned term.

## Statistical Indexing Properties of Databases

A fair amount of data exists on the statistical properties of databases (though we could benefit from quite a bit more), but these data have seldom been brought to bear on the discussion of indexing and access in IR system design. Here, three statistical features of databases that have implications for indexing and access to digital records will be discussed: Bradford's Law, vocabulary scalability, and the Resnikoff-Dolby 30:1 rule.

### *Bradford's Law*

The first point about statistical properties of databases is that there are very robust underlying patterns/distributions of vocabulary and index terms to be found in bibliographic and full-text databases. It is likely that many people involved with information organization and retrieval have not known or acted upon this fact.

Most information-related phenomena have been found to fall into a class of statistical distributions known as Zipfian distributions, named after George Zipf, who published data on the relationship of word frequency to word length (Zipf, 1949). In information science, Samuel Bradford developed what came to be known as Bradford's Law in describing the distribution of articles on topics through the journal literature (1948). Subsequently, Bradford's Law was found to appear in a variety

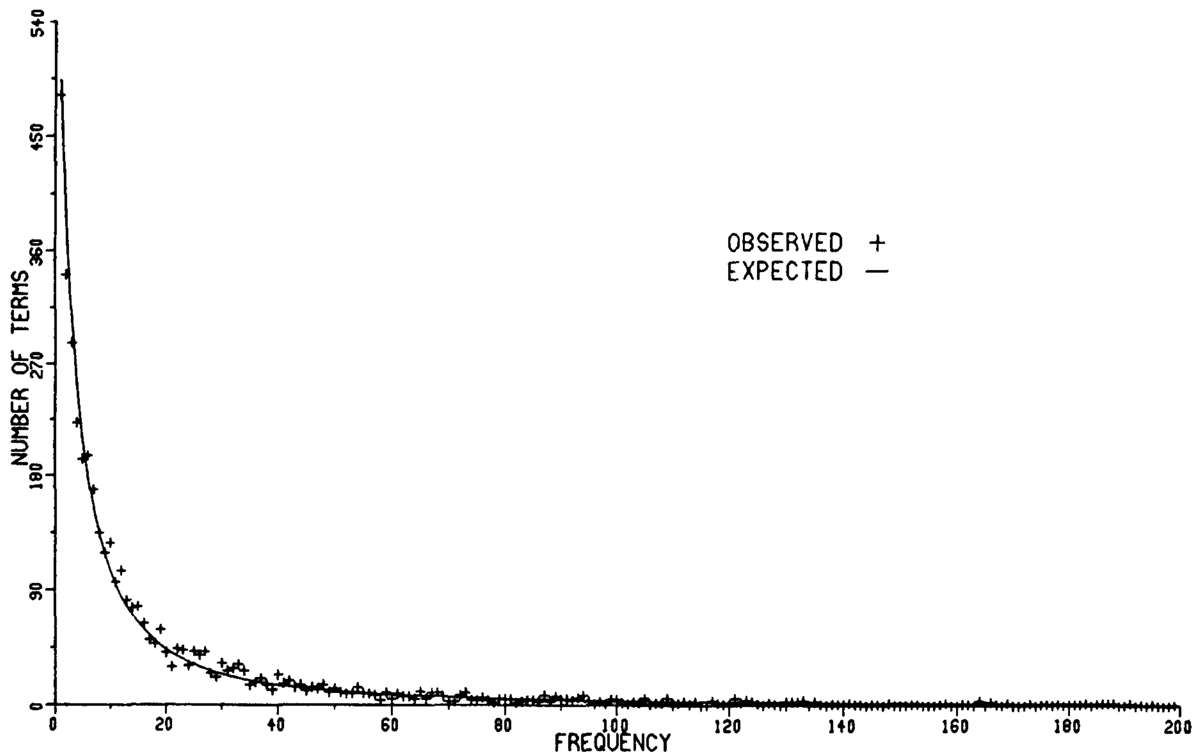


FIG. 2. A frequency-size Zipfian distribution. *Note.* This is Figure 2, reprinted from Nelson, M. J. and Tague, J. M. (1985). Split size-rank models for the distribution of index terms. *Journal of the American Society for Information Science, 36,* p. 291.

of other information-related phenomena (see Chen & Leimkuhler, 1986; Rousseau, 1994).

Other Zipfian formulations besides Zipf's and Bradford's are the distributions of Yule, Lotka, Pareto, and Price (Fedorowicz, 1982). Over the years, there has been extensive debate on the nature of the social processes that produce these distributions, and how best to represent them mathematically (e.g., Brookes, 1977; Chen & Leimkuhler, 1986; Qiu, 1990; Rousseau, 1994; Stewart, 1994). The intention here is not to justify any one Zipfian formulation—for our purposes they look similar—but rather to draw attention to the fact that information-related phenomena fall into this general class of distributions, as do many other distributions associated with human activity.

Zipfian distributions have a characteristic appearance which is quite different from the familiar bell-shaped normal distribution that is the usual focus of our statistics courses. Zipfian distributions are characterized by long tails and must often be presented on log paper in order to capture all the data produced in a research study. See Figure 2.

Zipfian distributions are presented as rank-frequency and frequency-size models (Nelson & Tague, 1985). For instance, a rank-frequency distribution of number of index terms assigned in a database displays the frequencies of application of the terms against their rank, with the most frequent being the highest rank (i.e., first). A frequency-size model shows the number of terms assigned in a database at each frequency level. In either case, there is generally a small number of terms with a very high frequency, and a large number of terms with a very low frequency. In Nelson and Tague's data, for example, approximately 1% of the terms were in the high-frequency portion of the distribution (p. 286). At the other end, the great majority of terms in a typical database might have been applied just once or twice.

These distributions are quite robust and defy efforts of thesaurus designers and indexers to thwart them. For instance, librarians have historically used some subdivisions in the *Library of Congress Subject Headings* to break down large sets of records indexed under the popular high-end terms into smaller sets. For example, "U.S.—History," an enormously popular term, would get another layer of subdivisions to produce headings like "U.S.—History—Revolution, 1775–83," and "U.S.—History—Civil War, 1861–65," and the like. However, the self-similar nature of this type of data indicates that the distribution of the subdivisions will *again* fall out into a Bradford form within the set of all headings beginning with "U.S.—History" (R. Rousseau, personal communication, 1996). Thus some of the combined headings will be very popular and some will have few or one application each, for a mini-Bradford distribution within the larger Bradford distribution.

In fact, Zipfian distributions are so common in information-related phenomena that there would be more surprise in finding that such data did *not* express a Zipfian pattern. To show how far afield this pattern can go, Brookes (1977) reports data from a 3-day-long closed conference with 50 people in attendance (pp. 207ff). One of the attending members kept a record of who contributed how many times each to the conference discussion. The results fell into a Bradford pattern, similar to what we have all experienced in school, where some students in class speak very frequently and many hardly at all.

To better understand the implications of these distributions for information retrieval, it may be helpful to visualize the Bradford distribution in a different way. When Bradford originally developed his data, he studied the rates at which articles relevant to a certain subject area appeared in journals in those areas. (His two test areas were applied geophysics and lubrication.) He identified all journals that published more than a certain number of articles in the test area per year, as well as in other ranges of descending frequency. He wrote:

> If scientific journals are arranged in order of decreasing productivity of articles on a given subject, they may be divided into a nucleus of periodicals more particularly devoted to the subject and several groups or zones containing the same number of articles as the nucleus, when the numbers of periodicals in the nucleus and succeeding zones will be as $1{:}n{:}n^2$. (Bradford, 1948, p. 116)

In principle, there could be any number of zones, with the number of articles in each zone being the total number of articles divided by the number of zones. In his empirical data, however, Bradford identified just three zones, and three will be used here for simplicity's sake. Bradford found in his empirical data that the value of "$n$" was roughly 5. Suppose, then, that someone doing an in-depth search on a topic finds that four core journals contain fully one-third of all the relevant articles found. If the value for $n$ is 5, then $4 \times 5 = 20$ journals will, among them, contain another third of all the relevant articles found. Finally, the last third will be the most scattered of all, being spread out over $4 \times 5^2 = 100$ journals. See Figure 3.

Figure 3 shows that one could find a great many relevant articles on a topic nicely concentrated in a few core journals. But finding the rest of the relevant articles involves an increasingly more extensive search, as the average yield of articles per additional journal examined becomes smaller and smaller the farther out, i.e., the more remotely from the core topic one goes. The Bradford data thus tell us that desired material is neither perfectly concentrated in one place, as we might hope, nor is it completely randomly distributed either, as we might fear.

If we assume that Bradford-type distributions of numbers of relevant hits per term in a database could also be arrayed as in Figure 3, then there are some predictable consequences in searching that can be anticipated. When searching with core, "right-on" terms, many relevant hits are retrieved, a high precision result. If, however, some high percentage of all the relevant articles are to be found "farther out," under still related but somewhat more remotely connected terms, then the percentage of good hits per added term goes down, the more remotely related the search term is.

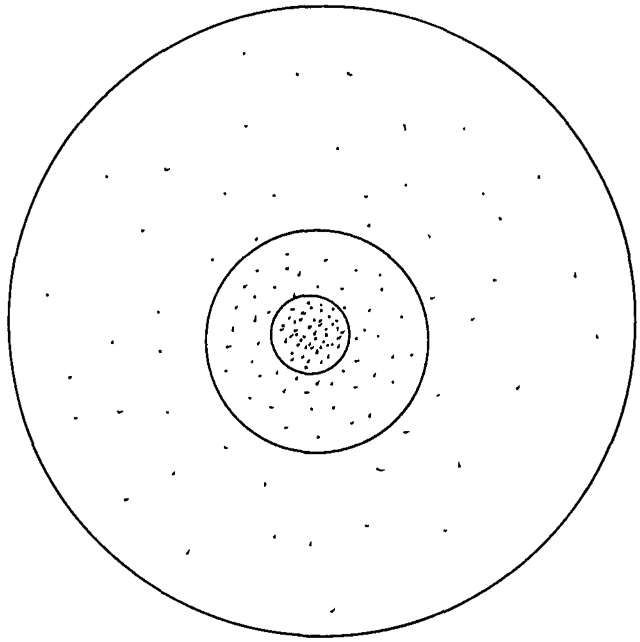

FIG. 3. Bradford scatter of hits across three zones. *Note.* Hits are sparser in sources in outer zones, therefore more sources must be covered—hence larger territory in each successive zone. Each zone contains the same total number of hits.

The pattern in Figure 3 would thus reflect the classic recall/precision trade-off. The better the recall, the lower the precision, because, as one incorporates more and more remotely related (OR'd) terms in the query, fewer and fewer relevant hits are found. To be confident in finding *all* relevant items, one would have to go very far afield, scooping up large numbers of records, with more and more remotely related terms, in order to find the occasional relevant item. The new system user who may never have thought about problems of retrieval may assume that a simple query will retrieve all relevant records; the experienced searcher knows that one must go very far afield to catch all relevant materials, and must tolerate increasingly many irrelevant records the farther out one goes.

Bradford's work was very important for special librarians to know about. It demonstrated that an organization that specializes in a certain subject area can readily purchase a good percentage of all relevant materials in its area by buying a small set of core journal subscriptions. But the higher the percentage of relevant materials the library seeks to capture through additional purchases, the smaller the

payoff per journal added. The same can be said, of course, of the recall/precision trade-off in IR.

So, what are the implications of this discussion for indexing digital resources? First, we can expect that even the most beautifully designed system of indexing—whether human or automatic—will produce Zipf-type distributions. If a system has a 10,000-term thesaurus, and a 100,000-record database, and each record has just one term applied to it, each term would then have 10 hits, if the distribution of term applications across the database were perfectly even. The research on these distributions, however, suggests that the distribution will be highly skewed, with some terms applied thousands of times, and a large number of the 10,000 terms applied just once.

Further, there is some evidence (Nelson, 1988) that frequencies of term applications in a database are significantly correlated with frequency of terms in queries, i.e., terms popular in indexing are also popular in information needs expressed as queries. So, those terms with lots of hits under them will also be used frequently among the queries coming into the database. This question needs much more research, however.

Finally, even in the best-designed database, users will usually retrieve a number of irrelevant records, and the more thorough they are in trying to scoop up every last relevant record by extending their search formulation with additional term variants, the greater the proportion of irrelevant records they will encounter. However, as the "core" terms will probably retrieve a relatively small percentage of the relevant records (certainly under half, in most cases), they must nonetheless tolerate sifting through lots of irrelevant records in order to find the relevant ones. It is the purpose of human indexing and classification to *improve* this situation, to pull more records into that core than would otherwise appear there, but it should be understood that even the best human indexing is not likely to defeat the underlying Zipfian patterns.

Indexers and system designers have to find a way to *work with* these database traits, rather than thinking they can, through good terminology design, produce a database full of terms, each of which falls in some ideal range of 10 to 50 hits. The latter, desirable as it may be, is not likely.

How then can we design systems to satisfy users, given this robust configuration of database term applications? I do not have any miracle solutions for it, but as long as these patterns are not recognized or understood, we will not even know how to begin to improve retrieval for users.

### *Vocabulary Scalability*

Scalability in the development of digital resources is one of the issues currently being discussed (Lynch & Garcia-Molina, 1995). Scalability deals with questions of "scaling up" systems from smaller to larger. Can a system that functions well at a certain size also function well when it is 10 or 100 times larger? This is an important question in information system design, because there is ample reason to believe that there are problems with scaling in this field. Small systems often do not work well when database size expands. New and different forms of organization and retrieval have to be introduced to maintain effectiveness.

Evidence of the need to change with growth in size can be detected in the longer history of information organization and access. In the 19th century, Charles Cutter developed alphabetical subject indexing ("subject headings") to make growing book collections more accessible. In the 1950s, when the conceptually broader subject headings began producing too many hits, especially in journal and report collections, "concept" indexing, which forms the basis of most current Boolean-searchable databases, was developed. When manual implementation of concept indexing bogged down on collections of even modest size, computer database searching was developed in the 1970s for faster searching.

Now that databases are exploding into the tens of millions in size, it can be guessed that new responses will again be needed. The purpose in this section is to draw attention to just one statistical feature that bears on scalability in indexing and/or full-text term matching as a means of retrieval from digital resources—the limitations of human vocabulary.

It is harder than it might at first seem to estimate the vocabulary size of the average person. Should the words one can produce be counted, or the much larger set one can recognize? Are proper noun names of products, people's names, etc., counted, or only "regular" lower-case words? How does one count the many morphological variations language creates for many words (e.g., compute, computes, computing, computed)? Here is Pinker (1994) on this question:

> The most sophisticated estimate comes from the psychologists William Nagy and Richard Anderson. They began with a list of 227,553 different words. Of these, 45,453 were simple roots and stems. Of the remaining 182,100 derivatives and compounds, they estimated that all but 42,080 could be understood in context by someone who knew their components. Thus there were a total of 44,453 + 42,080 = 88,533 listeme words. By sampling from this list and testing the sample, Nagy and Anderson estimated that an average American high school graduate knows 45,000 words. (p. 150)

This estimate did not include proper names, numbers, foreign words, and other such terms.

Not all of the 45,000 words would be suitable for searching for information. On the other hand, college graduates might know more words, including some that are technical

terms useful for searching and not in the general vocabulary. So, just to make a general estimate, let us take the closest round-number figure, 50,000, as the vocabulary size of an average information seeker on the Internet or in a digital collection.

A recent study found that 63% of the search terms employed by scholar searchers of databases were single-word terms (Siegfried, Bates, & Wilde, 1993, p. 282). (Solomon, 1993, also found that children overwhelmingly used single-word search terms.) So if users mostly employ single-word terms, let us, for the moment deal with just single-word indexing terms. Let us further assume that we have a million-item database—small by today's standards. We did arithmetic like this in the previous section—with 50,000 core terms available for searching by the user, and if the frequency in the database for each term were precisely even across the database, then each index term would have to have been applied to exactly 20 records (1,000,000/50,000). Such multiple uses of each term cannot be avoided when the size of the database is so much larger than the size of the indexing vocabulary.

But that assumes that only one term is applied per record. In fact, it is common for anywhere from 10 to 40 terms to be applied per record in human-indexed bibliographic databases, and, of course, indefinitely many—easily hundreds or more—words are indexed in each record in a full-text databases. Let us suppose our little million-item bibliographic database has an average of 20 index terms per record. Then, if the distribution of terms across the database were again exactly level, *each term would have 400 hits.*

This is a very problematic hit rate—it is far more than most people want. But we have only just begun. With natural-language indexing of the abstracts, let us suppose there are 100 indexable words per record, then if the application of words across the million-item database is level, there will be 2,000 hits for every indexable term used in the database. With digital libraries and the Internet, where tens of millions of records are currently involved, 20,000 or 50,000 or 100,000 hits per word can easily become *the minimum* on the assumption of level application of terms across the database.

There is no avoiding these statistics. As a database grows, the number of words a human being knows does not grow correspondingly. Consequently, the average number of hits grows instead.

It can be expected from the Bradford data, however, that the actual number of hits will be highly skewed, so some will have tens of thousands of hits and other terms very few. Further, if the popular terms are also popular in queries, per Nelson's data (1988), then a high proportion of all the queries will be pulling up that small set of terms with the huge numbers of hits.

To be sure, a searcher is not limited in most cases to single-word terms. The use of implicit or explicit Boolean logic in a query makes possible the combination of two or more words and usually drastically reduces the number of retrievals. Indeed, based on my own experience, I would estimate that a study on databases in the 1-million to 5-million item range would find that the best starting point for a search is to use two words or two concepts; that is, such a combination will most often produce sets of tolerable size. One-word searches produce too many high-hit searches and three-word (or concept) searches produce too many zero or low-hit searches. It remains to be seen whether three-word queries are optimal for searches on mega-databases of tens of millions of records.

Other problems arise at this point, however. There are astronomically many different combinations possible for even just two- and three-word phrases. If every one of our searcher's 50,000 words can be combined with every other one, then there will be just under 2.5 billion possible combinations for the two-word combinations alone. As we learned in an earlier section, varying the search terms by just morphological changes (not even included in that 50,000 figure)—"violence" to "violent," for example, produced substantial changes in the retrieved set, and varying by synonymous or near-synonymous terms often produced results that were entirely different.

So, by using more than one term, we succeed in reducing the retrieved set, but in ways that are quite unpredictable in relation to the question of which set of records will most please the searcher. How can I know a priori whether I should use "violent" or "violence" to find the hits that would most please me, if I could somehow look at all the records and determine which ones I would like? What if I think of "freedom of speech" and "cyberspace" and do not think of the other 15 combinations discussed earlier for this topic? Brilliant ranking algorithms for the retrieved set will have no impact on the many relevant unretrieved records.

The distribution of frequency of use of two-word combinations will probably fall out into a Bradford distribution too, so there will be the usual problem of some combinations with very many hits and many combinations with few or no hits. The "perfect 30-item search" that we all want from databases—actually, say, a range of 10–50 items—will likely occupy a small part of the entire distribution of hits in a database of almost any size, with most of the match rates being substantially larger or smaller.

For browsing purposes, the current indexing along with the ability to go in any of thousands of directions—sometimes with just the change of a character or two in a search statement (and, of course, also by following up hypertextual links)—makes the World Wide Web entertaining and sometimes informative. But active, directed searching for even a moderately focused interest, has yet to be provided for effectively, for data collections of the size available on the Net. The time has clearly arrived to develop a next round of innovations as information stores scale up another explosive level.

The general assumption has been that the needed innovations will be all automatic. Certainly, the solution will require substantial automated production, but what has been discussed so far in this article surely indicates that some role for human intervention is needed too. Algorithms based on current designs, or even enhancements of current designs, cannot deal with the many issues raised here.

### *Resnikoff-Dolby 30:1 Rule*

Earlier, in discussing Bradford-type distributions in information-related phenomena, we had to contemplate the possibility that our many individual actions and decisions as information system designers, and information providers and users, however motivated by individual choice and need, nonetheless fall out into remarkably persistent and robust statistical patterns. In examining the Resnikoff-Dolby 30:1 rule, we will discover an even more striking and hard-to-credit statistical pattern, which is nonetheless backed up by considerable data.

Under a grant from the Department of Health, Education, and Welfare, Howard Resnikoff and James Dolby researched the statistical properties of information stores and access mechanisms to those stores (Dolby & Resnikoff, 1971; Resnikoff & Dolby, 1972). Again and again, they found values in the range of 28.5:1 to 30:1 as the ratio of the size of one access level to another. For mathematical reasons, they used $K = 29.55$ as the likely true figure for their constant, but they and I will use 30 for simplicity's sake in most cases. They found from their data:

- A book title is $\frac{1}{30}$ the length of a table of contents in characters, on average (Resnikoff & Dolby, 1972, p. 10).
- A table of contents is $\frac{1}{30}$ the length of a back of the book index, on average (p. 10).
- A back of the book index is $\frac{1}{30}$ the length of the text of a book, on average (p. 10).
- An abstract is $\frac{1}{30}$ the length of the technical paper it represents, on average (p. 10).
- Card catalogs had one guide card for every 30 cards on average. Average number of cards per tray was $30^2$, or about 900 (p. 10).
- Based on a sample of over 3,000 four-year college classes, average class size was 29.3 (p. 22).
- In a test computer programming language they studied, the number of assembly language instructions needed to implement higher-level generic instructions averaged 30.3 (p. 96).

All these results suggest that human beings process information in such a way as to move through levels of access that operate in 30:1 ratios. Resnikoff and Dolby did not use this term, but I think a good name for it would be: *Information Transfer Exchange Ratio.* Something about these size relationships is natural and comfortable for human beings to absorb and process information. Consequently, the pattern shows up over and over again.

TABLE 1. Resnikoff and Dolby's "ranges in level structure."*

| Level | Type | No. of volumes Minimum | Mean = $K^n$ | Maximum |
|---|---|---|---|---|
| 1 | Encyclopedia | 5 | 30 | 161 |
| 2 | Personal | 161 | 873 | 4,747 |
| 3 | Juniorcollege | 4,747 | 25,803 | 140,266 |
| 4 | University | 140,266 | 762,483 | 4,144,851 |
| | | | | 122,480,276 |
| 5 | National | 4,144,851 | 22,531,361 | |

* Their Table 2, Resnikoff & Dolby (1972, p. 12). Used by permission.

But Resnikoff and Dolby's research did not stop there. They found that book superstructures, as well, tended to cluster around multiples of 30. Drawing on library statistical data, they found that junior college libraries clustered around $29.55^3$ volumes, or 25,803; university libraries clustered around $29.55^4$, or 762,483. The next level up, $29.55^5$, or 22,531,361, represented what they called a national library (Library of Congress) level. We would now add bibliographic utilities, such as OCLC, and, of course, Web browsers. With the spectacular growth of the Internet, we may soon expect to need to be able to provide access to resources in the $29.55^6$ range, or, in the 660-million range.

Of course, not all libraries or books fit the above averages. Libraries are constantly growing, and figures could be expected not to rest strictly on average points. But Resnikoff and Dolby did find that the data clustered fairly tightly around these means (1972, p. 92). The largest item at one level seldom exceeded the smallest item at another (p. 90).

Resnikoff and Dolby (1972) suggested that the line between two access levels should be drawn in the following manner: "Mathematically, the natural way to define a boundary between two values on an exponentially increasing scale is to compute the geometric mean of the two values" (p. 12). Their table listing these range values (p. 12), is adapted as Table 1.

Resnikoff and Dolby (1972) bring mathematical and physical arguments to bear in support of why this pattern appears, which will not be reproduced here. As they note:

> The access model presented in this chapter is not restricted to the book and its subsystems and super systems. There is considerable evidence that it reflects universal properties of information stored in written English form, and, in a slightly generalized version, may be still more broadly applicable to the analysis and modeling of other types of information systems such as those associated with the modalities of sensory perception. These wide ranging and difficult issues cannot be examined here in a serious way; moreover, we do not yet have sufficient data upon which a definitive report can be based. (p. 93)

To demonstrate results with such wide-ranging significance would indeed require vast amounts of data, more than one funded study could possibly produce. But Resnikoff and Dolby do bring forth a great deal of data from a wide variety of environments—more than enough to demonstrate that this is a fertile and very promising direction for research to go. (See also Dolby & Resnikoff, 1971. Later, Resnikoff, 1989, pursued his ideas in a biophysical context, rather than in library and information science applications.) While there are hundreds, if not more, papers pursuing questions related to Bradford distributions, this fascinating area of research has languished.[1]

Indeed, one may ask why these promising data have not been followed up more in the years since. When I have presented them, there is a common reaction that these relationships sound too formulaic or "gimmicky," that there could not possibly be such remarkable regularities underlying the organization of information resources. There has also been a strong shift in the years since from physical interpretations of information to constructivist views, which see each event or experience as uniquely constructed by the person doing the experiencing. I would argue that there are nonetheless regularities underlying our "unique" experiences, that while the individual construction of a situation generally cannot be predicted, there are often strong statistical patterns underlying human behavior over a class of events. We will find that there is no fundamental contradiction in the physical/statistical and the social/constructivist views; only a failure to accept that they can, and do, operate at the same time.

Equipped with the ideas in the Resnikoff-Dolby research, it is possible to see how certain results in the research literature—done independently of Resnikoff and Dolby's work—in fact, fit nicely within the framework of their model of information. Wiberley, Daugherty, and Danowski did successive studies on user persistence in displaying online catalog postings, first on a first-generation online catalog (1990), and later on a second-generation online catalog (1995). In other words, they studied how many postings users actually examined when presented with a hit rate of number of postings found when doing a search in an online catalog. They summarize the results of both studies in the abstract of the second (1995):

> Expert opinion and one study of users of a first-generation online catalog have suggested that users normally display no more than 30 to 35 postings. . . . Analysis of transaction logs from the second-generation system revealed that partially persistent users typically displayed 28 postings, but that overloaded users did not outnumber totally persistent users until postings retrieved exceeded 200. The findings suggest that given sufficient resources, designers should still consider 30 to 35 postings typical persistence, but the findings also justify treating 100 or 200 postings as a common threshold of overload. (p. 247)

Note the parallels between this data and that in Table 1. Resnikoff and Dolby's level 1 shows a mean of 30 items, with a maximum of 161. In the Wiberley, Daugherty, and Danowski research, people feel comfortable looking at 30 items, and will, if necessary go up to "100 or 200" (ideally, 161?), but beyond that figure go into overload.

Another parallel appears in the research on library catalogs. Time and time again, when asked, catalog users say they want more information, an abstract or contents list, on the catalog record (Cochrane & Markey, 1983). Resnikoff and Dolby discuss at some length questions of the role the catalog plays, in terms of levels of access, to the collection. The argument will not be made here in detail, but both bodies of data converge on the idea that people need another 30:1 layer of access between the catalog entry and the book itself. This need shows up in the request for an abstract or summary, presumably 30 times as long in text as the average book title or subject heading. The Resnikoff and Dolby research also clearly needs to be related to the research on menu hierarchies in the human computer interaction literature (e.g., Fisher, Yungkurth, & Moss, 1990; Jacko & Salvendy, 1996).

Finally, at many points in their analysis, Resnikoff and Dolby (1972) work with the lognormal distribution: "These examples and others too numerous to report here prompt us to speculate that the occurrence of the lognormal distribution is fundamental to *all* human information processing activities" (p. 97). That distribution is one of the common candidates for modeling Zipfian distributions (see Stewart, 1994). Those with more mathematical expertise than I will surely be able to link these two bodies of data in a larger framework of analysis.

It may yet be found that human information processing characteristics are the underlying driving force behind many other seemingly random or unpredictable statistical characteristics of the information universe. If so, it may be discovered that the two seemingly disjoint approaches taken so far in this article, the "Human Factors" and the "Statistical Indexing Properties of Databases," are, in fact, very closely linked and both derive from common sources.

## Domain-Specific Indexing

Research emanating from two areas of study strongly indicates that attention to subject domain in the design of

[1] For those unable to access the ERIC technical report, White, Bates, and Wilson (1992) have a brief summary and discussion on p. 271. White says there that in writing my paper "The Fallacy of the Perfect 30-Item Online Search" (Bates, 1984), I did not know of Dolby's work. That is not strictly true. I took a class from Dolby when I was a graduate student at the University of California at Berkeley, where he mentioned the 30:1 ratio and some of the instances where the ratio had been found. The class preceded the publication of the technical report, however, and I did not know until White wrote his chapter that Dolby's figure was based on extensive data and had been published through ERIC. Had I known of the work, I would have addressed it in my own analysis of book indexes, tables of contents, etc. (in Bates, 1986b).

content indexing and access has potentially high payoff in improved results for users. Research in computational linguistics and automatic translation has increasingly been looking to sub-domains of language as more productive venues for improving effectiveness in application of automatic techniques (Grishman & Kittredge, 1986; Kittredge & Lehrberger, 1982). The value of sublanguage analysis is beginning to be recognized and studied in information science as well (Haas & He, 1993; Liddy, Jorgensen, Sibert, & Yu, 1993; Losee & Haas, 1995).

Research in library and information science on domain vocabularies for indexing and retrieval has likewise demonstrated the importance of domain in the provision of access to resources. Stephen Wiberley has demonstrated that the character of the humanities vocabulary does not match the conventional assumptions that such vocabulary is fuzzy and hard to pin down (Wiberley, 1983, 1988). Nor is the vocabulary similar in character to natural and social scientific vocabulary, according to a study carried out at the Getty Information Institute (Bates, Wilde, & Siegfried, 1993; Bates, 1996b). That study found several classes of vocabulary types used in the humanities literature that were virtually non-existent in the science literature. Tibbo (1994) surveys the distinct issues associated with humanities indexing, Swift, Winn, and Bramer (1979) draw attention to unique characteristics of social science vocabulary, and Pejtersen (1994) applies domain analysis to fiction classification.

Weinberg (1988) also makes a strong case for a whole different class of access to information, viz., *aspect,* or *comment,* as distinct from the usual "aboutness" that index-descriptive terms are intended to provide. This sort of information would be particularly valuable for the humanities and the "softer" social sciences. Yet it is the characteristics of scientific and engineering vocabulary that have driven most of the theory of indexer thesaurus development since World War II.

Likewise, Mavis Molto, who was interested in developing automated techniques for genealogical literature, found that the distribution of term types in that literature is very different from general text at large, and that means of optimizing retrieval in such text may be quite different from those assumed previously (Molto, 1993). These results suggest dramatic differences in the character of vocabulary and information needs across the various intellectual domains. Techniques that work well in one may not work well in another.

There are other reasons, as well, for suspecting that treatment of the various intellectual domains as though all vocabularies are equal is an ineffectual and sub-optimal approach. Research in the sociology of science (Meadows, 1974; Merton, 1973) and scholarly communication (Bakewell, 1988; Bates, 1994, 1996a; Morton & Price, 1986) demonstrate a number of fundamental differences in how scholars and scientists conceptualize their research problems, in how paradigms do and do not function in various domains, in how social processes and relationships among researchers function differently in one domain as compared to another, and on and on. We should rather assume that these many differences *do* make a difference in means to optimize access to recorded materials than that they do not—at least until such time as we have much more research in hand on the effectiveness of various access techniques and vocabulary types.

Recently, Hjørland and Albrechtsen (1995) made an extensive and cogent argument for the value of domain analysis in information science. Their argument will not be adumbrated here, except to quote them briefly:

> The domain-analytic paradigm in information science (IS) states that the best way to understand information in IS is to study the knowledge-domains as thought or discourse communities, which are parts of society's division of labor. Knowledge organization, structure, co-operation patterns, language and communication forms, information systems, and relevance criteria are reflections of the objects of the work of these communities and their role in society. The individual person's psychology, knowledge, information needs, and subjective relevance criteria should be seen in this perspective. (p. 400)

Much could be written and hypothesized, well beyond the scope of this article, on how domain factors might affect retrieval. The goal here, however, is rather simply to establish, in general, the importance and value of paying attention to domain as a way to improve retrieval. So, the final argument to be brought forth is a practical, real-world one, a description of a database that uses domain analysis and has met the test of the marketplace.

BIOSIS, the database of biological periodicals, is just one of several that might serve as examples, but illustrates particularly well the points being made in this section. The database contains over 10 million records and adds over half a million records per year (*BIOSIS Search Guide,* 1995, p. A-3), so it represents a very large body of information with, as we shall see, substantial value added in the handling of its indexing, both human and automatic.

The developers of BIOSIS studied the literature of the field and the needs of their users, and created a resource that serves many of the distinctive needs of the biological research community. Articles are coded for whether they describe a new species, for example, a likely important question for many biologists searching the database.

But the analysis of the literature of the field goes much deeper than that. In the end, the database makes available two broad classes of information—*taxonomic* (descriptions of species and broader classes of flora and fauna) and *general subject* information. These are, in turn, broken out in ways likely to be of particular value in the study of biology.

The Biosystematic Index provides searchable codes for taxonomic levels discussed in articles above the genus-

**CC07006 ANIMAL COMMUNICATION**

***Frequencies*** Major (6760) Minor (540)

***Applications*** This code retrieves studies on nonverbal human communication and on physical and biochemical animal communication pathways and patterns.

***Examples*** Studies on • echolocation • pheromones • displays • animal sounds • infant noises

***Strategy Recommendations***

- For comparative studies of animal communication, use this code with appropriate keywords and the *General and Comparative Behavior* code CC07002.
- For human speech or verbal communication, use the *Deafness, Speech and Hearing* code CC20008.

***Coverage Note*** This code was introduced in 1971.

FIG. 4. BIOSIS concept code for animal communication. *Note.* Concept code taken from *BIOSIS Search Guide* (1995, p. F-11). Reprinted by permission of BIOSIS.

species level. At the species level, species can be searched by the user as free text on the subject-related fields of the record. Retrieval is improved, however, by the indexers at BIOSIS, who add species names to the basic bibliographic description of the article in every case that they are mentioned in the text of the article, but not in the title or abstract.

General subject topics are, likewise, approached at both broad and specific levels. Broad, popular concepts, the sort that recur frequently in the biological literature, have been identified and carefully described and analyzed for both indexers and searchers. Such broad concepts include things like animal distribution, wildlife management, and reproduction. Because these are both common and broad, searching on these terms alone or with a species name could lead to very large and wildly varying retrieval sets, depending on which particular terms for the concept had been used by the article's writer, the indexer, or the searcher. Because of the importance of these concepts (they are clearly at the upper end of the Bradford distribution), it has paid off for BIOSIS to analyze these terms closely and define them precisely for both indexers and searchers. These terms are called "concept codes." For example, Figure 4 displays the entry in the BIOSIS search manual for the concept code for Animal Communication.

On the other hand, less common subject terms, those that are more specific and narrowly applicable, and often at the low end of the Bradford distribution, do not receive this elaborate analysis. They are, instead, not even indexed in any formal way; they are just made free-text searchable. BIOSIS helps the searcher improve results, however, by providing a search manual containing the more popular of these free-text terms. For the less popular terms, or terms newly appearing since the last manual, the searcher knows to use them anyway, because the list of terms in the manual is explicitly stated to include only the more popular terms, and not all legitimate ones, as is usual with thesauri.

BIOSIS has thus cleverly worked simultaneously with the character of the domain and its practitioners' information needs, the character of the terminology important in the domain, and the statistical character of the literature in the domain, to produce a database that enables fast, high-recall, high-precision retrieval.

## Discussion and Conclusions

Having reviewed these various issues, we come to the next natural question: How can these things we have learned be applied to improve IR system design for the new digital information world?

In the previous section, the BIOSIS database of biological literature was brought forth as an example demonstrating a good understanding of and use of domain factors in the design of indexing for very large databases. Special indexes and retrieval capabilities probed precisely the kinds of information of most interest and use to biologists. Further, as was noted there, BIOSIS also effectively uses the Bradfordian statistical properties of its material, by indexing deeply and carefully the core concepts in biology that recur frequently in the literature, and therefore constitute the high-frequency end of the Bradford curve.

The BIOSIS database constitutes one solution that integrates both database statistical factors and subject domain factors in its design. It demonstrates that the seemingly intractable Bradfordian characteristics—the massively skewed distributions that typically characterize meaningful collections of information—need not be defeated in order to effectuate good information retrieval. Rather, these statistical characteristics can be *worked with,* the system design ingeniously set up to invest high indexing effort in the high-demand end of the curve, and low effort, largely automatic, in the low-demand end.

There are no doubt other such inspired mixes of human and system effort possible throughout the information world, once the underlying factors are understood. One area that remains to be studied much more is that of the Resnikoff-Dolby 30:1 rule. If, as the evidence to date suggests, people often like to move down through a hierarchy of some sort as they access information, various kinds of hierarchical arrangements could be developed with layers sized in a 30:1 ratio.

Many systems currently are being developed with the structure of classical hierarchical (family-tree style) classification schemes. In library and information science, however, such classifications are frequently seen to be less effective than classifications based on faceted principles, developed some decades ago by S.R. Ranganathan (1963) but only now coming into their full fruition. Examples of faceted approaches can be seen in the Getty *Art & Architecture Thesaurus* and in the Predicasts business database. Faceting provides a flexibility and power much in excess of the limited character of conventional hierarchical schemes—not unlike the advance represented by relational databases in the database world. (A brief, simplified comparison of the two types of classification scheme is provided in Bates, 1988. See Rowley, 1992, or Vickery, 1968, for

more thorough introductions.) Using this and other indexing techniques, in combination with a sensitivity to database and domain factors, can produce great retrieval power. Whether and how these techniques can be used to good effect in multi-million-item collections remains to be seen.

However, of all the issues raised in this article, those in the first section, on human factors, have been least attended to in system design (see also Chen & Dhar, 1990). Little research in information science has examined the various folk classification issues raised in this article—whether people spontaneously use terms that fall into groupings fitting the characteristics of folk classifications, whether there are Roschian "basic level" terms (see also Brown, 1995), and so on.

Many design changes need to be made in information systems in response to the research on the great variety of terms generated by people for a given topic. Until we fully accept that reality, and build information retrieval systems that intelligently incorporate it, systems will continue to under-perform for users. It is interesting to note, however, that two different information systems, designed for two radically different environments, independently came up with solutions involving the collecting and clustering together of variant terms around core concepts.

The Los Angeles Department of Water and Power, with more than 10,000 employees, needed to solve a thesaurus problem for their multi-million item records management database. The records would be input into the system by clerks who often had only a high school education and no concept of vocabulary control, and retrieved by any and everyone in the organization, up to and including the General Manager. The content range of the vocabulary needed for retrieval was very wide—engineering, real estate, customer relations, construction, names of organizations, officers, etc.

The solution proposed (Bates, 1990), which was subsequently implemented, was to use highly skilled human labor to create and update continuously a cluster thesaurus of closely related term variants, based on a close monitoring of database contents and relevant technical thesauri. (In this way, only one person, the thesaurus developer, needed to have lexicographical expertise, rather than expecting every clerk in the organization to develop it.) When a searcher used any one of the terms, the variants could be called up too. The searcher was then given the option of searching on all or selected terms from the cluster. The person searching on "transmission line right of way" might think no variants are needed until the cluster appears with a dozen or more other ways the concept has been described in the database (some examples: transmission line easements, T/L Right of Way, T/L ROW).

Meanwhile, at the Getty Information Institute, the rarefied world of art historical scholarship was likewise being served by a database developed to cluster term variants—in this case, various forms of artists' names, in the *Union List of Artist Names* (1994). Just as it was impractical at the Water and Power Department to get all the department's memo and report writers to use the same word for the same concept, it is also undesirable to try to search on an artist's name under just one form. Artists have frequently been known by both vernacular and Latin names, by different names in different languages, and by nicknames and formal names. They may be referred to in all these forms in the literature of interest to a researcher, and the researcher may not initially be aware of all the forms that have been used. Consequently, provision of a database containing all those forms not only aids in retrieval effectiveness, but is an important scholarly resource in its own right.

The Getty Information Institute has been experimenting with making this and two other thesaural databases, *Art & Architecture Thesaurus* (Peterson, 1994), and *Thesaurus of Geographic Names,* available on the Internet. As of this writing, two of the three are combined in a online thesaurus named "a.k.a." which can be used during an experimental period by people searching the Getty bibliographic databases via the World Wide Web (http://www.ahip.getty.edu/aka/). Searching on the artist "Titian" yielded the names of five different individuals who have been known by this name, as well as all the variants found on each individual's name. One of these five, Tiziano Vecellio, has been known by at least 46 different names, according to the *Union List of Artist Names.* Examples include the following: Detiano, Titsiaen, Tizzani, Vecelli, and Ziano. This term-variant database provides a nice example of ways in which online access can be improved in real time for Internet or other online searchers.

A number of other experimental systems and approaches are currently being developed to improve vocabulary for users for search and retrieval as well (Chen & Lynch, 1992; Chen & Ng, 1995; Chen, Yim, Fye, & Schatz, 1995; Johnson & Cochrane, 1995; Jones et al., 1995; Kristensen, 1993; Larson, McDonough, O'Leary, Kuntz, & Moon, R., 1996; Schatz, Johnson, Cochrane, & Chen, 1996).

Both the Water and Power and the Getty systems rely on a valuable aspect of human cognition that is unfortunately much ignored in the information system design world: People can *recognize* information they need very much more easily than they can *recall* it. The average person will recall (think up) only a fraction of the range of terms that are used to represent a concept or name, but can take in a screen full of variants in an instant, and make a quick decision about desired terms for a given search. Most current information systems require that the searcher generate and input everything wanted. People could manage more powerful searches quickly if an initial submitted term or topic yielded a screen full of term possibilities, related subjects, or classifications for them to see and choose from (see also Bates, 1986a).

These and other experiments in improved access are made possible if *indexing* is separated from *access* in the design of an information system. By separating these two, the user front-end can be designed around the distinctive traits and evolutionary adaptations of human information

processing, while the internal indexing describing the document may be different. Changes in our developing understanding of human cognition in information-seeking situations, and changes in vocabulary can relatively quickly be accommodated in a user-oriented front-end, without requiring the re-indexing of giant databases.

In sum, the pieces of knowledge reviewed here have yet to be unified into a single grand model. Much research remains to be done. However, there are tantalizing hints that seemingly dramatically different features presented herein—such as the human and statistical factors—may indeed ultimately be given a unified formulation, once research and theory development have progressed far enough.

From what is already known, however, it can be seen that we need to design for the real characteristics of human behavior as people have information needs and confront information systems to try to meet those needs. We know, further, that robust underlying statistical properties of databases demand that we find ingenious ways to work with these statistical patterns, rather than deny or ignore them. Finally, indexing and information system design needs to proceed within the context of a deep understanding of the character of the intellectual domain, its culture, and its research questions. Any design lacking understanding of these human, database, and domain elements will almost certainly sub-optimize information retrieval from digital resources and the Internet.

## References


Bakewell, E. (1988). *Object, image, inquiry: The art historian at work.* Santa Monica, CA: Getty Art History Information Program.

Bates, M. J. (1977). System meets user: Problems in matching subject search terms. *Information Processing & Management, 13,* 367–375.

Bates, M. J. (1984). The fallacy of the perfect thirty-item online search. *RQ, 24,* 43–50.

Bates, M. J. (1986a). Subject access in online catalogs: A design model. *Journal of the American Society for Information Science, 37,* 357–376.

Bates, M. J. (1986b). What is a reference book? A theoretical and empirical analysis. *RQ, 26,* 37–57.

Bates, M. J. (1988). How to use controlled vocabularies more effectively in online searching. *Online, 12,* 45–56.

Bates, M. J. (1989a). The design of browsing and berrypicking techniques for the online search interface. *Online Review, 13,* 407–424.

Bates, M. J. (1989b). Rethinking subject cataloging in the online environment. *Library Resources & Technical Services, 33,* 400–412.

Bates, M. J. (1990). Design for a subject search interface and online thesaurus for a very large records management database. *Proceedings of the 53rd ASIS Annual Meeting* (pp. 20–28). Medford, NJ: Learned Information.

Bates, M. J. (1994). The design of databases and other information resources for humanities scholars: The Getty online searching project report no. 4. *Online and CDROM Review, 18,* 331–340.

Bates, M. J. (1996a). Document familiarity, relevance, and Bradford's Law: The Getty online searching project report no. 5. *Information Processing & Management, 32,* 697–707.

Bates, M. J. (1996b). The Getty end-user online searching project in the humanities: Report no. 6: Overview and conclusions. *College & Research Libraries, 57,* 514–523.

Bates, M. J., Wilde, D. N., & Siegfried, S. (1993). An analysis of search terminology used by humanities scholars: The Getty online searching project report no. 1. *Library Quarterly, 63,* 1–39.

Belkin, N. J. (1982). ASK for information retrieval. Part 1: Background and theory. *Journal of Documentation, 38,* 61–71.

Belkin, N. J., Kantor, P., Fox, E. A., & Shaw, J. A. (1995). Combining the evidence of multiple query representations for information retrieval. *Information Processing & Management, 31,* 431–448.

Berlin, B. (1978). Ethnobiological classification. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 9–26), Hillsdale, NJ: Lawrence Erlbaum.

*BIOSIS search guide.* (1995). Philadelphia, PA: BIOSIS.

Blair, D. C. (1996). STAIRS redux: Thoughts on the STAIRS evaluation, ten years after. *Journal of the American Society for Information Science, 47,* 4–22.

Borgman, C. L. (1996). Why are online catalogs still hard to use? *Journal of the American Society for Information Science, 47,* 493–503.

Bradford, S. C. (1948). *Documentation.* London: Crosby Lockwood.

Brookes, B. C. (1977). Theory of the Bradford Law. *Journal of Documentation, 33,* 180–209.

Brooks, T. A. (1995). People, words, and perceptions: A phenomenological investigation of textuality. *Journal of the American Society for Information Science, 46,* 103–115.

Brown, C. H. (1984). *Language and living things: Uniformities in folk classification and naming.* New Brunswick, NJ: Rutgers University Press.

Brown, M. E. (1995). By any other name: Accounting for failure in the naming of subject categories. *Library & Information Science Research, 17,* 347–385.

Chen, H., & Dhar, V. (1990). User misconceptions of information retrieval systems. *International Journal of Man-Machine Studies, 32,* 673–692.

Chen, H., & Lynch, K. J. (1992). Automatic construction of networks of concepts characterizing document databases. *IEEE Transactions on Systems, Man, and Cybernetics, 22,* 885–902.

Chen, H., & Ng, T. (1995). An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): Symbolic branch-and-bound search vs. connectionist Hopfield net activation. *Journal of the American Society for Information Science, 46,* 348–369.

Chen, H., Yim, T., Fye, D., & Schatz, B. (1995). Automatic thesaurus generation for an electronic community system. *Journal of the American Society for Information Science, 46,* 175–193.

Chen, Y. S., & Leimkuhler, F. F. (1986). A relationship between Lotka's law, Bradford's Law, and Zipf's Law. *Journal of the American Society for Information Science, 37,* 307–314.

Cochrane, P. A., & Markey, K. (1983). Catalog use studies—before and after the introduction of online interactive catalogs: Impact on design for subject access. *Library and Information Science Research, 5,* 337–363.

Dolby, J. L., & Resnikoff, H. L. (1971). On the multiplicative structure of information storage and access systems. *Interfaces: The Bulletin of the Institute of Management Sciences, 1,* 23–30.

Eichman, T. L. (1978). The complex nature of opening reference questions. *RQ, 17,* 212–222.

Ellen, R. F., & Reason, D. (Eds.). (1979). *Classifications in their social context.* New York: Academic Press.

Ellis, D. (1996). The dilemma of measurement in information retrieval research. *Journal of the American Society for Information Science, 47,* 23–36.

Fedorowicz, J. (1982). The theoretical foundation of Zipf's Law and its application to the bibliographic database environment. *Journal of the American Society for Information Science, 33,* 285–293.

Fidel, R. (1994). User-centered indexing. *Journal of the American Society for Information Science, 45,* 572–576.

Fisher, D. L., Yungkurth, E. J., & Moss, S. M. (1990). Optimal menu hierarchy design—Syntax and semantics. *Human Factors, 32,* 665–683.

Furnas, G. W., Landauer, T. K., Dumais, S. T., & Gomez, L. M. (1983). Statistical semantics: Analysis of the potential performance of keyword information systems. *Bell System Technical Journal, 62,* 1753–1806.

Gomez, L. M., Lochbaum, C. C., & Landauer, T. K. (1990). All the right words: Finding what you want as a function of richness of indexing

vocabulary. *Journal of the American Society for Information Science, 41,* 547–559.
Green, R., & Bean, C. A. (1995). Topical relevance relationships. II. An exploratory study and preliminary typology. *Journal of the American Society for Information Science, 46,* 654–662.
Grishman, R., & Kittredge, R. (Eds.). (1986). *Analyzing language in restricted domains: Sublanguage description and processing.* Hillsdale, NJ: Lawrence Erlbaum.
Haas, S. W., & He, S. (1993). Toward the automatic indentification of sublanguage vocabulary. *Information Processing & Management, 29,* 721–732.
Harter, S. P. (1992). Psychological relevance and information science. *Journal of the American Society for Information Science, 43,* 602–615.
Hert, C. A. (1996). User goals on an online public access catalog. *Journal of the American Society for Information Science, 47,* 504–518.
Hjørland, B., & Albrechtsen, H. (1995). Toward a new horizon in information science: Domain-analysis. *Journal of the American Society for Information Science, 46,* 400–425.
Houston, J. E. (Ed.). (1995). *Thesaurus of ERIC Descriptors.* 13th ed. Phoenix, AZ: Oryx.
Jacko, J. A., & Salvendy, G. (1996). Hierarchical menu design—Breadth, depth, and task complexity. *Perceptual and Motor Skills, 82,* 1187–1201.
Johnson, E., & Cochrane, P. (1995). A hypertextual interface for a searcher's thesaurus. In *Proceedings of Digital Libraries '95,* Austin, TX, June 11–13, 1995 (pp. 77–86). College Station, TX: Hypermedia Research Laboratory.
Jones, S., Gatford, M., Robertson, S., Hancock-Beaulieu, M. et al. (1995). Interactive thesaurus navigation: Intelligence rules OK? *Journal of the American Society for Information Science, 46,* 52–59.
Kaske, N. K., & Sanders, N. P. (1980a). Evaluating the effectiveness of subject access: The view of the library patron. *Proceedings of the 43rd Annual ASIS Meeting, 17,* 323–325.
Kaske, N. K., & Sanders, N. P. (1980b). On-line subject access: The human side of the problem. *RQ, 20,* 52–58.
Kittredge, R., & Lehrberger, J. (1982). *Sublanguage: Studies of language in restricted semantic domains.* New York: Walter de Gruyter.
Klatt, M. J. (1994). An aid for total quality searching: Developing a hedge book. *Bulletin of the Medical Library Association, 82,* 438–441.
Knapp, S. D. (Ed.). (1993). *The contemporary thesaurus of social science terms and synonyms: A guide for natural language computer searching.* Phoenix, AZ: Oryx.
Kristensen, J. (1993). Expanding end-users' query statements for free text searching with a search-aid thesaurus. *Information Processing & Management, 29,* 733–744.
Kuhlthau, C. C. (1993). *Seeking meaning: A process approach to library and information services.* Norwood, NJ: Ablex.
Larson, R. R., McDonough, J., O'Leary, P., Kuntz, L., & Moon, R. (1996). Cheshire II: Designing a next-generation online catalog. *Journal of the American Society for Information Science, 47,* 555–567.
Leonard, L. E. (1977). *Inter-indexer consistency studies, 1954–1975: A review of the literature and summary of the study results.* Champaign-Urbana, IL: University of Illinois Graduate School of Library Science. Occasional papers No. 131.
Library of Congress. Cataloging Policy and Support Office. (1996). *Library of Congress Subject Headings* (19th ed.). Washington, DC: Cataloging Distribution Service, Library of Congress.
Library of Congress. Office for Subject Cataloging Policy. (1991–). *Subject cataloging manual. Subject headings* (4th ed.). Washington, DC: Cataloging Distribution Service, Library of Congress.
Liddy, E. D., Jorgensen, C. L., Sibert, E. E., & Yu, E. S. (1993). A sublanguage approach to natural language processing for an expert system. *Information Processing & Management, 29,* 633–645.
Lilley, O. L. (1954). Evaluation of the subject catalog. *American Documentation, 5,* 41–60.
Losee, R. M., & Haas, S. W. (1995). Sublanguage terms: Dictionaries, usage, and automatic classification. *Journal of the American Society for Information Science, 46,* 519–529.
Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development, 1,* 309–317.
Lynch, C., & Garcia-Molina, H. (1995). *Interoperability, scaling, and the digital libraries research agenda: A report on the May 18–19, 1995 IITA Digital Libraries Workshop, August 22, 1995* [Online]. Available: http://www.diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.html.
Lynch, M. J. (1978). Reference interviews in public libraries. *Library Quarterly, 48,* 119–142.
Markey, K. (1984a). Interindexer consistency tests: A literature review and report of a test of consistency in indexing visual materials. *Library & Information Science Research, 6,* 155–177.
Markey, K. (1984b). *Subject searching in library catalogs: Before and after the introduction of online catalogs.* Dublin, OH: OCLC Online Computer Library Center.
Meadows, A. J. (1974). *Communication in science.* London: Butterworths.
Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations.* Chicago: University of Chicago Press.
Molto, M. (1993). Improving full text search performance through textual analysis. *Information Processing & Management, 29,* 615–632.
Morton, H. C., & Price, A. J. (1986). The ACLS survey: Views on publications, computers, libraries. *Scholarly Communication, 5,* 1–16.
Nelson, M. J. (1988). Correlation of term usage and term indexing frequencies. *Information Processing & Management, 24,* 541–547.
Nelson, M. J., & Tague, J. M. (1985). Split size-rank models for the distribution of index terms. *Journal of the American Society for Information Science, 36,* 283–296.
Newport, E. L., & Bellugi, U. (1978). Linguistic expression of category levels in a visual-gestural language: A flower is a flower is a flower. In E. Rosch, & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 49–71), Hillsdale, NJ: Lawrence Erlbaum.
O'Connor, B. C. (1996). *Explorations in indexing and abstracting.* Englewood, CO: Libraries Unlimited.
Pejtersen, A. M. (1994). A framework for indexing and representation of information based on work domain analysis: A fiction classification example. In H. Albrechtsen, & S. Oernager (Eds.), *Knowledge organization and quality management* (pp. 161–172). Frankfurt am Main, Germany: Indeks Verlag.
Peterson, T. (1994). *Art & architecture thesaurus* (2nd ed.). New York: Oxford.
Pinker, S. (1994). *The language instinct.* New York: Harper Perennial.
Qiu, L. (1990). An empirical examination of the existing models for Bradford's Law. *Information Processing & Management, 26,* 655–672.
Ranganathan, S. R. (1963). *Colon classification: Basic classification* (6th ed.). New York: Asia Publishing House.
Raven, P. H., Berlin, B., & Breedlove, D. E. (1971). The origins of taxonomy. *Science, 174,* 1210–1213.
Resnikoff, H. L., & Dolby, J. L. (1972). *Access: A study of information storage and retrieval with emphasis on library information systems.* Washington, DC: U.S. Department of Health, Education, and Welfare, Office of Education, Bureau of Research. (ERIC Document Reproduction Service No. ED 060 921).
Rosch, E. (1978). Principles of categorization. In E. Rosch, & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Lawrence Erlbaum.
Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8,* 382–439.
Rousseau, R. (1994). Bradford curves. *Information Processing & Management, 30,* 267–277.
Rowley, J. (1992). *Organizing knowledge: An introduction to information retrieval* (2nd ed.). Brookfield, VT: Ashgate.

Saracevic, T., & Kantor, P. (1988). A study of information seeking and retrieving. III. Searchers, searches, and overlap. *Journal of the American Society for Information Science, 39,* 197–216.

Schatz, B. R., Johnson, E. H., Cochrane, P. A., & Chen, H. (1996). Interactive term suggestion for users of digital libraries: Using subject thesauri and co-occurrence lists for information retrieval. In *Proceedings of the 1st ACM International Conference on Digital Libraries* (pp. 126–133). New York: Association for Computing Machinery.

Siegfried, S., Bates, M. J., & Wilde, D. N. (1993). A profile of end-user searching behavior by humanities scholars: The Getty online searching project report no. 2. *Journal of the American Society for Information Science, 44,* 273–291.

Soergel, D. (1985). *Organizing information: Principles of database and retrieval systems.* Orlando, FL: Academic Press.

Solomon, P. (1993). Children's information retrieval behavior: A case analysis of an OPAC. *Journal of the American Society for Information Science, 44,* 245–264.

Stevens, M. E. (1965). *Automatic indexing: A state-of-the-art report.* Washington, DC: U.S.G.P.O. National Bureau of Standards Monograph 91.

Stewart, J. A. (1994). The Poisson-lognormal model for bibliometric/scientometric distributions. *Information Processing & Management, 30,* 239–251.

Swift, D. F., Winn, V. A., & Bramer, D. A. (1979). A sociological approach to the design of information systems. *Journal of the American Society for Information Science, 30,* 215–223.

Tibbo, H. R. (1994). Indexing for the humanities. *Journal of the American Society for Information Science, 45,* 607–619.

*Union list of artist names* (Version 1.0) [computer file]. (1994). New York: G. K. Hall/Macmillan.

Vickery, B. C. (1968). *Faceted classification: A guide to construction and use of special schemes.* London: Aslib.

Weinberg, B. H. (1988). Why indexing fails the researcher. *The Indexer, 16,* 3–6.

White, H. D., Bates, M. J., & Wilson, P. (1992). *For information specialists: Interpretations of reference and bibliographic work.* Norwood, NJ: Ablex.

Wiberley, S. E. (1983). Subject access in the humanities and the precision of the humanist's vocabulary. *Library Quarterly, 53,* 420–433.

Wiberley, S. E. (1988). Names in space and time: The indexing vocabulary of the humanities. *Library Quarterly, 58,* 1–28.

Wiberley, S. E., Daugherty, R. A., & Danowski, J. A. (1990). User persistence in scanning postings of a computer-driven information system: LCS. *Library and Information Science Research, 12,* 341–353.

Wiberley, S. E., Daugherty, R. A., & Danowski, J. A. (1995). User persistence in displaying online catalog postings: LUIS. *Library Resources & Technical Services, 39,* 247–264.

Wilson, P. (1968). *Two kinds of power: An essay on bibliographical control.* Berkeley, CA: University of California Press.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology.* Cambridge, MA: Addison-Wesley.

# Software Engineering as Seen through Its Research Literature: A Study in Co-Word Analysis

**Neal Coulter**
*Department of Computer Science and Engineering, Florida Atlantic University, Boca Raton, FL 33431.*
*E-mail: neal@cse.fau.edu*

**Ira Monarch and Suresh Konda**
*Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA 15213-3890.*
*E-mail: {iam}@sei.cmu.edu; {slk}@sei.cmu.edu*

**This empirical research demonstrates the effectiveness of content analysis to map the research literature of the software engineering discipline. The results suggest that certain research themes in software engineering have remained constant, but with changing thrusts. Other themes have arisen, matured, and then faded as major research topics, while still others seem transient or immature. Co-word analysis is the specific technique used. This methodology identifies associations among publication descriptors (indexing terms) from the ACM Computing Classification System and produces networks of descriptors that reveal these underlying patterns. This methodology is applicable to other domains with a supporting corpus of textual data. While this study utilizes index terms from a fixed taxonomy, that restriction is not inherent; the descriptors can be generated from the corpus. Hence, co-word analysis and the supporting software tools employed here can provide unique insights into any discipline's evolution.**

## Introduction

Software engineering is a term often used to describe programming-in-the-large activities. Yet, any precise empirical characterization of its conceptual contours and their evolution is lacking. Many professionals have discussed these issues (Buckley, 1993; Coulter & Dammann, 1994; Denning et al., 1989; Ford & Gibbs, 1989; Gibbs, 1989, 1991; Parnas, 1990, 1997; Shaw, 1990; Tucker, 1991). A special ACM/IEEE Computer Society task force is now commissioned to consider them (Boehm, 1994).

Received August 4, 1997; revised November 10, 1997; accepted November 10, 1997.



In this study, a large number of 1982–1994 publications—each represented by descriptor (or index) terms—are analyzed to determine themes and trends in software engineering research. We believe that empirically driven answers to the questions formulated will not only enrich previous and ongoing discussions, but will help in providing an effective basis for determining research, application, and curriculum targets in industry and academe.

Empirical studies require a carefully considered methodology and accompanying data sets. The methodology we have chosen is based on co-word analysis (Callon, Courtial, & Laville, 1991; Callon, Law, & Rip, 1986; Courtial, 1994; Courtial & Law, 1989; Law & Whittaker, 1992; Turner, Chartron, Laville, & Michelet, 1988; Whittaker, 1989). Co-word analysis reveals patterns and trends in technical discourse by measuring the association strengths of terms representative of relevant publications or other texts produced in a technical field. A main tenet of co-word analysis is that the identified patterns of representative term associations are maps of the conceptual space of technical fields, and that a series of such maps constructed for different time periods produces a trace of the changes in this conceptual space.

Co-word analysis is related to co-citation analysis (Small, 1973; Small & Griffith, 1974) and related less closely to bibliographic coupling (Kessler, 1963; Weinberg, 1974). Co-citation analysis provides a method of mapping the structure of a research field through pairs of documents that are jointly cited. See Savoy (1997) for a recent discussion of co-citation analysis and bibliographic coupling. See Garfield (1983) and Liu (1993) for general discussions of citation indexing, and its use in information retrieval and related matters.

Co-word analysis deals directly with sets of terms shared by documents instead of with shared citations. Therefore, it maps the pertinent literature directly from the interactions of key terms instead of from the interactions of citations. Future work should compare the two approaches, evaluate their respective strengths and weaknesses, and investigate whether they can be combined in productive ways. Together, they might enable a richer approach for studying and characterizing the evolution of disciplines, as part of the basis for targeting research and development funds, and modifying curricula.

We applied co-word analysis to a very large cross-section of published text (1982–1994) in the computing field, indexed by descriptors from the well-known ACM Computing Classification System (CCS).[1] This indexed text comes from the Association for Computing Machinery's (ACM) *Guide to Computing Literature* (GUIDE), which covers an ACM publications database. Through professional indexers, GUIDE annually covers over 20,000 publications by descriptors from the CCS.[2]

CCS is a carefully designed and maintained taxonomy; it was created in 1982 and has been updated four times (Coulter, 1998). Because CCS classifies publications over the breadth of computing, it allows us to investigate trends and the position of software engineering in the larger computing context.

The fully study is quite extensive. Here we summarize only our major findings. Complete details of this research are available in a Carnegie Mellon University/Software Engineering Institute technical report (Coulter, Monarch, Konda, & Carr, 1996).

## The Data

GUIDE reviews and indexes a large number of published documents across the spectrum of computing. Documents reviewed generally include books, book chapters, journals, proceedings, trade magazines, and other applied sources, and occasionally other media such as videotaped material. The range and currency of publications assure reasonably even coverage and currency of software engineering topics. For the list of publication titles received, see the November 1996, *Computing Reviews* (CR, 1996). In addition, GUIDE indexes many proceedings and proceedings articles.

The major CCS categories are:

A—General Literature
B—Hardware
C—Computer Systems Organization
D—Software
E—Data
F—Theory of Computation
G—Mathematics of Computing
H—Information Systems
I—Computing Methodologies
J—Computer Applications
K—Computing Milieux

The full CCS is described in the January 1997 issue of *Computing Reviews* (CR, 1997).

### *CCS Category D.2—Software Engineering*

A complete description of the D.2—Software Engineering section of the taxonomy follows. Superscripts indicate that a descriptor is new beginning with the year indicated; all others are from the 1982 implementation.

D.2 SOFTWARE ENGINEERING
D.2.0 General
  Protection mechanisms
  Standards
D.2.1 Requirements/Specifications
  Languages
  Methodologies
  Tools
D.2.2 Tools and Techniques
  Computer-aided software engineering (CASE)$^{91}$
  Decision table
  Flow charts
  Modules and interfaces
  Petri nets$^{91}$
  Programmer workbench
  Software libraries
  Structured programming
  Top-down programming
  User interfaces
D.2.3 Coding
  Pretty printers
  Program editors
  Reentrant code
  Standards
D.2.4 Program Verification
  Assertion checkers
  Correctness proofs
  Reliability
D.2.5 Testing and Debugging
  Code inspections and walk-throughs$^{91}$
  Debugging aids
  Diagnostics
  Dumps
  Error handling and recovery
  Symbolic execution
  Test data generators
  Tracing
D.2.6 Programming Environments

[1] Until 1996, the computing Classification System (CCS) was called the Computing Reviews Classification System (CRCS).

[2] Descriptors selected from CCS are distinguished from keywords freely chosen by the author. Only CCS descriptors were used in this study. We believe it is useful to study a fixed system that imposes a common nomenclature across all computing. Professional indexers experienced in using the CCS assure standard application of that taxonomy. It would be interesting to investigate what the differences are among descriptors selected by professional indexers, by the free selection of authors, or by automated derivation from the texts themselves. Law and Whittaker (1992) and Whittaker (1989) have addressed some of the issues involved.

Interactive[87]
D.2.7 Distribution and Maintenance
Corrections
Documentation
Enhancement
Extensibility
Portability
Restructuring
Version control
D.2.8 Metrics
Complexity measures
Performance measures
Software science
D.2.9 Management
Copyrights
Cost estimation
Life cycle
Productivity
Programming teams
Software configuration management
Software quality assurance
Time estimation[91]
D.2.10 Design[87]
Methodologies[87]
Representation[87]
D.2.m Miscellaneous
Rapid prototyping[83]
Reusable software[83]

Documents are almost always indexed by multiple CCS descriptors. Even though there are up to four CCS levels, documents are classified at any appropriate level. CCS does not include names of systems and languages (Unix, Ada, Windows, etc.) as explicit descriptors; instead, they are called *implicit descriptors* and can be used by indexers as needed. As we will see, their inclusion is common and often significant.

We obtained descriptors for all documents indexed in GUIDE that had at least one descriptor in the D.2—Software Engineering category. This selection allows us to examine interactions of software engineering descriptors with other descriptors in CCS. We could have refined this study by selecting more specific CCS descriptors [such as how Software Engineering (D.2) interacts with programming Techniques (D.1), for example]. However, this study focuses on the larger question of how software engineering interacts with computing as a whole; i.e., on the interactions of software engineering with all other descriptors of the CCS.[3]

The data we received reflects the March 1995 update to the GUIDE database. We received the following numbers of indexed documents for the years 1982–1994 shown in Table 1.

The total is 16,691 documents. As is evident, the number of documents was small until 1986. The 16,691 documents were indexed by a total of 57,727 descriptors (a mean of 3.46 per document).

[3] We show the corresponding CCS node after a descriptor when context is needed.

TABLE 1. Distribution of documents by year.

| Year | No. of documents | Year | No. of documents |
|---|---|---|---|
| 1982 | 81 | 1989 | 1,928 |
| 1983 | 33 | 1990 | 1,914 |
| 1984 | 211 | 1991 | 1,738 |
| 1985 | 367 | 1992 | 2,016 |
| 1986 | 1,027 | 1993 | 2,159 |
| 1987 | 1,479 | 1994 | 1,612 |
| 1988 | 2,329 | Total | 16,691 |

For analysis, we grouped the data for the years 1982–1986, 1987–1990, 1991–1994. This grouping separates the sparse years 1982–1986 from the others, it gives approximately equal numbers of documents in the latter two periods, and it provides breaks when CCS was updated so we do not confuse new descriptors across periods. Data for documents, descriptors, and descriptors/documents ratios for the time periods are shown in Table 2.

## The Metric and the Algorithm

### *Details of the Metric Used*

Co-word analysis enables the structuring of data at various levels of analysis: (1) As networks of links and nodes (nodes hold descriptors from CCS; links connect nodes, thereby forming networks); (2) as distributions of interacting networks; and (3) as transformation of networks over time periods. These structures and changing relationships provide a basis for tracing the evolution of software engineering.

Co-word analysis reduces a large space of related descriptors to multiple related smaller spaces that are easier to comprehend but are also indicative of actual partitions of interrelated concepts in the literature under consideration. This analysis requires an association measure and an algorithm for searching through a descriptor space. The analysis is designed to identify areas of strong focus that interrelate. This scheme allows us to construct a mosaic of software engineering topics.

Metrics for co-word analysis have been studied extensively (Callon et al., 1986, 1991; Courtial & Law, 1989; Law & Whittaker, 1992; Whittaker, 1989). Two descriptors, $i$ and $j$, co-occur if they are used together in the classification of a single document. Take a corpus consisting of $N$ documents. Each document is indexed by a set of unique descriptors that can occur in multiple documents. Let $c_k$ be the number of times $k$ is used for indexing documents in the corpus. Let $c_{ij}$ be the number of co-occurrences of descriptors $i$ and $j$ (the number of documents indexed by both descriptors).

Different measures of association have been proposed. The *inclusion index* $I_{ij}$ provides a hierarchy of association metric (essentially a conditional probability) through the function:

TABLE 2. Documents and descriptors per time period.

| Time period | Documents | Descriptors | Descriptor/ document ratio |
|---|---|---|---|
| 1982–1986 | 1,646 | 5,645 | 3.43 |
| 1987–1990 | 7,650 | 28,471 | 3.72 |
| 1991–1994 | 7,395 | 23,611 | 3.19 |

$$I_{ij} = \frac{c_{ij}}{\min(c_i, cj)}.$$

$I_{ij}$ is not symmetrical (or bidirectional) and tends to highlight mainly the central poles in a domain and depict their relations with descriptors that occur less frequently (Callon et al., 1986). Moreover, the inclusion hierarchies discovered using this metric do not cohere with common semantic relations like part–whole and subcategorization.

The basic metric used for this study is *Strength* $S_{ij}$. The Strength $S$ of association between descriptors $i$ and $j$ is given by the expression:

$$S_{ij} = \frac{c_{ij}^2}{c_i \cdot c_j}, \; 0 \le S \le 1.$$

This metric does not impose the conceptual inclusion property of the inclusion index $I_{ij}$, though it does provide an intuitive measure of the strength of association between terms, indicating only that there is some semantic relationship or other. The metric is easier to understand and utilize in the production and interpretation of term association maps than is the other metric. It allows associations of both major and minor descriptors and is symmetrical in their relationships (Callon et al., 1991). $S$ can be used as the basis for several complementary measures of interactions of descriptors and descriptor networks in a unified manner.

Two descriptors that appear many times in isolation, but only a few times together, will yield a lower $S$ value than two descriptors that appear relatively less often alone, but have a higher ratio of co-occurrences are included in or excluded from networks based on having relatively high $S$ values. A network consists of nodes (descriptors) connected by links. Each node must be linked to at least one other node in a network.

The co-word algorithm uses two passes through the data to produce pair-wise connections of descriptors in networks (see Fig. 1, which is described below). Pass-1 builds networks that can identify areas of strong focus; Pass-2 can identify descriptors that associate in more than one network, and thereby indicate pervasive issues. This pattern of networks yields a mosaic of the data being analyzed.

The first pass (Pass-1) generates the primary associations among descriptors; these descriptors are called *internal nodes* and the corresponding links are called *internal links.* A second pass (Pass-2) generates links between Pass-1 nodes across networks, thereby forming associations among completed networks. Pass-2 nodes and links are called *external* ones.

Figure 1 illustrates this process for a 1991–1994 network. This figure displays the network connections as a map.[4] This network, named User Interfaces,[5] is the first one created by the co-word algorithm for 1991–1994 data. Pass-1 links and nodes are represented by thick lines connecting thick boxes, respectively. Pass-2 nodes are in thin boxes, while Pass-2 links are shown as thin lines connecting Pass-1 and Pass-2 nodes. The links are numbered in the order formed. The notation *N-x* beneath each Pass-2 node indicates its Pass-1 network number, $x$.

Without some minimum constraints, descriptors appearing infrequently but almost always together could dominate networks; hence a minimum co-occurrence $c_{ij}$ value is required to generate a link. At the same time, some maps can become cluttered due to an excessive number of legitimate links (but of generally decreasing $S$ values); hence, restrictions on numbers of nodes and links are sometimes required to help discover major partitions of concepts. However, many networks are limited only by the number of qualifying nodes, as we will observe.

For the time periods of 1987–1990 and 1991–1994, a descriptor co-occurrence of 15 was required for linking; for 1982–1986, the co-occurrence cutoff was set at five to accommodate the lesser volume of data. For all time periods, the number of links and nodes in each network, both Pass-1 and Pass-2, was set at 24 links and 20 nodes. For these values the co-word algorithm generated 15, 16, and 11 networks, respectively, for the periods 1982–1986, 1987–1990, and 1991–1994. Table 3 summarizes these values.

### *Details of the Co-Word Algorithm Used*

During Pass-1, the link that has the highest strength is selected first. These linked nodes become the starting points for the first network. Other links and their corresponding nodes are then determined breadth-first. All nodes contained in the resulting Pass-1 network are removed from consideration for inclusion in subsequent Pass-1 networks. The next network then starts with the link of highest $S$ value of the remaining links (i.e., ones not containing nodes from any previous network).

The second pass (Pass-2) is designed to seek further associations among descriptors found in Pass-1. During Pass-2, networks are extended by the addition of Pass-2 links. To be a candidate for inclusion in Pass-2, both nodes (descriptors) of a Pass-2 link must be in some Pass-1 networks. A Pass-2 link connects a Pass-1 node in a given network to a node that had occurred as a Pass-1 node in another network (but is represented in the given network as a Pass-2 node).

[4] These were originally called Leximappes (Turner et al., 1988).

[5] Names of networks are underlined for clarity.

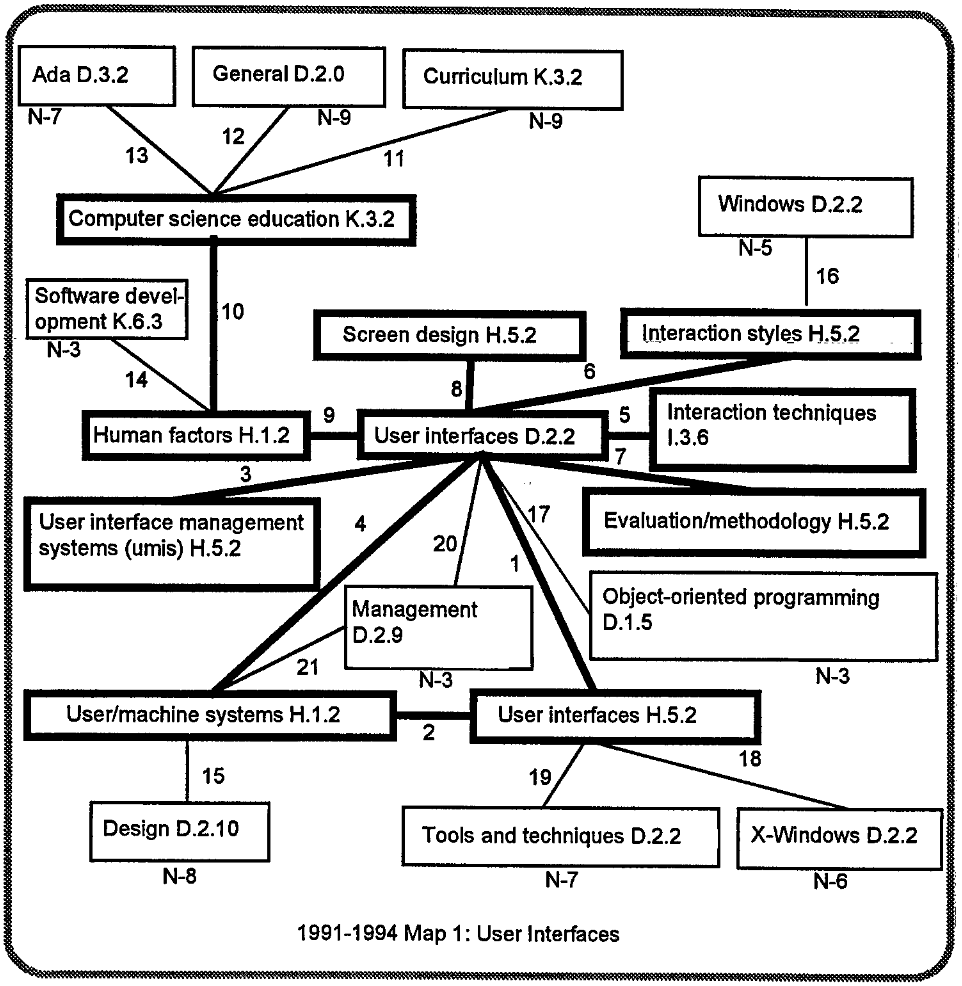


FIG. 1. Pass-1 and Pass-2 nodes and links.

As in Pass-1, candidate links are included in Pass-2 based on their strengths and co-occurrence counts. The order of Pass-2 links is by descending values for qualifying links. A node can appear in only one Pass-1 network, but can appear in more than one Pass-2 link.

The steps in the analysis procedure are:

1. Select a minimum for the number of co-occurrences, $c_{ij}$, for descriptors $i$ and $j$.
2. Select maxima for the number of Pass-1 links and nodes;
3. Select maxima for the total (Pass-1 and Pass-2) links and nodes;
4. Start Pass-1;
5. Generate the highest $S$ value from all possible descriptors to begin a Pass-1 network;
6. From that link, form other links in a breadth-first manner, until no more links are possible due to the co-occurrence minima or to Pass-1 link or node maxima. Remove all incorporated descriptors from the list of subsequent available Pass-1 descriptors;
7. Repeat Steps 5 and 6 until all Pass-1 networks are formed; i.e., until no two remaining descriptors co-occur frequently enough to begin a network;
8. Begin Pass-2;
9. Restore all Pass-1 descriptors to the list of available descriptors;

TABLE 3. Parameters and resulting networks.

| Time period | Minimum co-occurrence | Maximum nodes | Maximum links | Networks generated |
|---|---|---|---|---|
| 1982–1986 | 5 | 20 | 24 | 15 |
| 1987–1990 | 15 | 20 | 24 | 16 |
| 1991–1994 | 15 | 20 | 24 | 11 |

10. Starting with the first Pass-1 network, generate all links to Pass-1 nodes in that network with any Pass-1 nodes having at least the minimal co-occurrences in descending order of $S$ value; stop when no remaining descriptors meet co-occurrence minima or when total node or link maxima are met. Do not remove an descriptors from the available list;
11. Repeat Step 10 for each succeeding Pass-1 network.

A maximum number of Pass-1 networks can be specified in cases where an excessive number of networks will be generated otherwise; this restriction was not necessary here.

## Comments on Network Parameter Selection

Link and node constraints mostly determine how networks will be generated in concert with the corresponding co-occurrence minimum. If the co-occurrence minimum is too high, few links may be formed; if it is too low, an excessive number of links may result. In the former case, subspecialities in a field may not emerge; in the latter case, more representative and well-connected themes and problem spaces will be harder to detect amidst the noise of less representative and less well-connected ones. We experimented with many different sets of parameters during our research to identify values that yield a detailed, yet coherent, set of networks.

The parameters for 1982–1986 were chosen somewhat arbitrarily because of the small amount of data. We attempted to establish a baseline for comparison with following generations. The primary point of contention was the co-occurrence value of five. It is somewhat higher in proportion to the number of documents and descriptors than the value of 15 for succeeding generations. However, we feel the number of networks and super networks generated is justifiable—at least with respect to our current capabilities of handling large numbers of different patterns of co-occurrence. In setting co-occurrence values for the 1987–1990 and 1991–1994 generations, the proper values could be determined at least two ways: As a function of the ratios of indexed items or the ratio of the number of descriptors. We used the former. Because the numbers of items for the generations were almost equal (7,650 and 7,395), we set the co-occurrences the same. However, the numbers of descriptors were sufficiently different (28,471 and 23,611) to question if the co-occurrence for 1991–1994 should be lower than for 1987–1990. To test this hypothesis, we set the 1991–1994 co-occurrence at 13.

This change still resulted in 11 networks. Some networks were different, but only on the fringes. The central themes remained the same. More links and nodes were realized with the lower co-occurrence value (16 and 19%, respectively), as would be expected. Many of these new links and nodes were formed through additional connections of already existing nodes in the same and in other maps existing at the higher co-occurrence level. Additionally, one isolated network with only two nodes was absorbed by a larger network at the 13 co-occurrence level, while a new, isolated network with three nodes and two links emerged.

So, while the link, node, and co-occurrence parameters effectively control the generation of networks, small changes in their values appear to affect only marginal links. Of course, additional and subsequent data can affect the generation of core themes without changes in parameters, which is the intent of co-word analysis.

## The Software Tools Used

Co-word analysis is an active research area at Carnegie Mellon University's Software Engineering Institute[6] (SEI). A suite of software tools developed at SEI called CAIR (Content Analysis and Information Retrieval) performs all of the calculations described here based on an algorithm also developed at SEI/CMU. CAIR operates on documents represented by index terms (as in GUIDE data) and it operates on free-text documents, using both noun phrase and stem words. Recently, CAIR was employed to analyze a free-text database of best engineering practices to extract which best practices might be applicable in software engineering (Coulter & Monarch, 1996a). CAIR is used extensively at SEI to analyze various collections of free text to assist in understanding and categorization of software engineering management practices (Monarch, 1994, 1996; Monarch & Gluch, 1995).

CAIR works with the probabilistic information retrieval system INQUERY[7] to allow users to query a database for specific documents whose key term associations are revealed through lexical maps. A graphical user interface (gui) builds term association maps and connects them to a document database enabling the formation of searches by selecting linked nodes or clustered document icons. The gui can be used to depict the semantic distance of documents with respect to one another, and to the nodes on the term association map (not shown here). It enables a selected map to be ranked against other maps for similarity and dissimilarity. It can be used to create a reduced document set derived from an ongoing search from which a new set of term association maps can be derived. It also provides other convenient features such as automating layout of networks, and modifying them for enhanced presentation.

## Related Work

The use of co-word analysis to develop and/or refine the taxonomy of a field is thus well-established. Several

[6] The Software Engineering Institute is a federally funded research and development center sponsored by the U.S. Department of Defense and Operated by Carnegie Mellon University.

[7] INQUERY is licensed from Applied Computing Systems Institute of Massachusetts, Inc. (ACSIOM).

strong European research teams continue to apply these techniques, sometimes in combination with other techniques (Courtial, Cahlik, & Callon, 1994; Turner, Lelu, & Georgel, 1994; Turner & Rojouan, 1991; Zitt, 1991; Zitt & Bassecoulard, 1994). In North America (aside from the examples already mentioned), co-word analysis has been integrated in knowledge level support systems for scholarly communities (Gaines & Shaw, 1994). Prototypes have been built that combine term association maps with knowledge acquisition and knowledge representation techniques. The intention is to build detailed formal knowledge structures to provide a framework for knowledge expression, interchange, and collaborative development. So far, this work has produced fairly detailed scenarios, thereby setting an agenda for scholarly communication by identifying the computational resources available, and describing the social-technical infrastructure (Gaines, 1993) needed for long-term collaborative development of classification and knowledge representation systems, in specific technical domains.

The Illinois Digital Library (IDL) project is also concerned with generating concept maps for information retrieval on the basis of term co-occurrence analysis (Schatz & Chen, 1996; Schatz, Johnson, Cochrane, & Chen, 1996a; Schatz et al., 1996b). In these studies, co-occurrence analysis is much different from what is being described here as co-word analysis. The former does not produce a collection of autonomous but interrelated subnetworks or maps as does co-word analysis. Instead, it produces co-occurrence lists for key terms which, collected together, are called concept graphs. These are used to bridge vocabulary differences among groups seeking the same information, but who use different terms to describe the same concept because of their varying backgrounds (called the *vocabulary problem*—Chen, 1994). Concept graphs, unlike term association maps, are not used to describe an identifiable research or application area in a technical field. Concept graphs are not used to depict the structure of subject and problem areas, their relative coherence and interdependencies, their evolution, or their comparison in these regards to networks of other fields, as the present study has done in comparing the evolving networks of computer science and software engineering. While the IDL project wants to evolve to conceptual spaces for semantic retrieval and vocabulary switching across different technical domains, it proceeds on the assumption that solving the vocabulary problem involves *automatically* mapping a community library's specialized terms into the corresponding terms of other libraries being searched. Intersecting co-occurrence lists from different domains provides an approach to interactive term suggestion across community libraries.

The work described here proceeds from a different—though, in some ways, related—perspective. From this perspective, the library of the near future will not only be a repository of documents to be searched, but a basis for mapping the emergence, cross-fertilization, and evolution of various technical domains. Mapping the evolutionary dynamics of science and technology will enhance information retrieval technology, and other information and knowledge management activities, but will also provide a richer guide for researchers, as well as decision and policy makers. The aim is not just to improve interactive term suggestion for automated or semi-automated query elaboration on the basis of co-occurrence lists, but to discover the conceptual network structures that underlie conceptual translation and transformation. This requires careful attention to how knowledge is organized and structured for people, as well as machines (Mylopoulos, Borgida, & Yu, 1997).

In our view, the vocabulary problem needs to be seen from a knowledge representation perspective that combines intellectual collaboration and computational environments. The vocabulary problem is a translation problem. Translation between disciplines (just as translation between natural languages) is not fully automatable. It depends on knowledge structures constructed collaboratively with the help of knowledge construction tools (Monarch 1994, 1996; Monarch & Nirenburg, 1993). As with Gaines and Shaw (1994) above, co-word analysis is only part of the process for building detailed formal knowledge structures, such as cross-disciplinary classification systems. Moreover, as with Gaines (1993), Levy (1995), and Monarch et al. (1997), building these knowledge structures requires socio-technical infrastructures to be established and maintained. Building the interspace that the IDL project wants to build is not a purely computational problem. The Computing Classification System (CCS) is a case in point.

## Network Names and Structures

We named the maps in an attempt to summarize their main thrusts. This is not a precise activity. In some cases maps had a lone, highly connected descriptor; in others, two (and sometimes three) descriptors exhibited high connectivity. In the latter cases, we used hyphenated names. We always used descriptors contained in nodes. Generally, we used descriptor(s) from the node(s) with the most connections, giving greater weight to Pass-1 nodes in close calls. We used a D.2 descriptor as a name if possible.

For reference in the following sections, the primary network names for each time period are given in Table 4. Networks are listed in the order generated by co-word analysis algorithms for each time period. The identification number beside the entries will be utilized in subsequent discussions.

Shown with each network are the total numbers of nodes and links (under the N-L heading). Notice the span of network sizes generated by the algorithm from the data. While some networks are relatively large, others approach

TABLE 4. Network names and numbers.

| 1982–1986 | N-L | 1987–1990 | N-L | 1991–1994 | N-L |
|---|---|---|---|---|---|
| 1—Software management—Ada | 18–24 | 1—Geometrical problems and computations | 6–6 | 1—User interfaces | 20–21 |
| 2—Logic programming | 2–1 | 2—Correctness proofs—Languages | 17–22 | 2—Petri nets | 2–1 |
| 3—User interfaces | 17–24 | 3—Logic programming | 3–2 | 3—Software development—Object-oriented programming | 20–23 |
| 4—Standards | 2–1 | 4—Requirements/specifications—Methodologies | 16–24 | 4—Software libraries—C++—Microsoft windows | 17–24 |
| 5—Tools and techniques—Structured programming—Pascal | 20–23 | 5—User/machine systems | 15–24 | 5—Windows | 17–17 |
| 6—Software development | 20–22 | 6—Methodologies—Software development | 20–23 | 6—X-windows | 8–7 |
| 7—Software libraries | 3–2 | 7—Standards | 3–2 | 7—Tools and techniques—CASE—Systems analysis and design | 17–24 |
| 8—Testing and debugging—Correctness proofs | 20–23 | 8—Structured programming | 4–3 | 8—Requirements/specifications | 16–23 |
| 9—Reliability | 6–6 | 9—Applications and expert systems—Tools and techniques | 19–24 | 9—General | 5–5 |
| 10—Program editors | 3–2 | 10—Concurrent programming–Ada | 18–24 | 10—Concurrent programming | 4–3 |
| 11—Requirements/specifications—Systems analysis and design | 17–24 | 11—Computer-aided design | 4–3 | 11—Metrics | 5–5 |
| 12—Modules and interfaces | 2–1 | 12—Error handling and recovery | 2–1 | | |
| 13—Real-time systems | 3–2 | 13—Distribution and maintenance | 4–3 | | |
| 14—Abstract data types | 3–2 | 14—Software configuration management | 6–5 | | |
| 15—Metrics | 12–13 | 15—Reusable software | 16–24 | | |
| | | 16—Software management—Design | 11–20 | | |

or reach minimal formation, which is two nodes with one link.

## Types of Networks and Their Interactions

There are essentially three types of networks: Principal, secondary, and isolated. principal networks are connected to one or more (secondary) networks. Secondary networks generally are linked to principal networks through a relatively high number of external links in the principal networks. Isolated networks have an absence (or low intensity) of links with other networks.

Isolated networks often have links with high $S$ values, usually accompanied by low co-occurrence $c_{ij}$ values. While isolated networks are easy to recognize, principal and secondary networks may not be. Therefore, we will define and operationalize terms that characterize these functionalities.

We defined *density* as the mean of the Pass-1 $S$ values of a network; *centrality* is defined as the square root of the sum of the squares of the Pass-2 $S$ values of a network, in order to distinguish among relatively close values. Density represents the internal strength of a network, while centrality represents a network's position in strength of interaction with other networks.[8]

Plots of centrality and density for each of the time periods are shown in Figures 2 3, 4.[9] The origin of these figures is the median of the represent axis values (the horizontal axis represents centrality; the vertical axis represents density).

Not surprisingly, most networks with strong centrality scores also show relatively high node and link counts.

[8] These terms are accepted ones in co-word analysis literature. We recognize that density and centrality have other domain-specific connotations—say, in statistics. Alternative choices include *cohesion* and *coupling*, but these already have meanings in software engineering literature.

[9] Figures 2, 3, and 4 are not to precise scale; relative positions are represented.

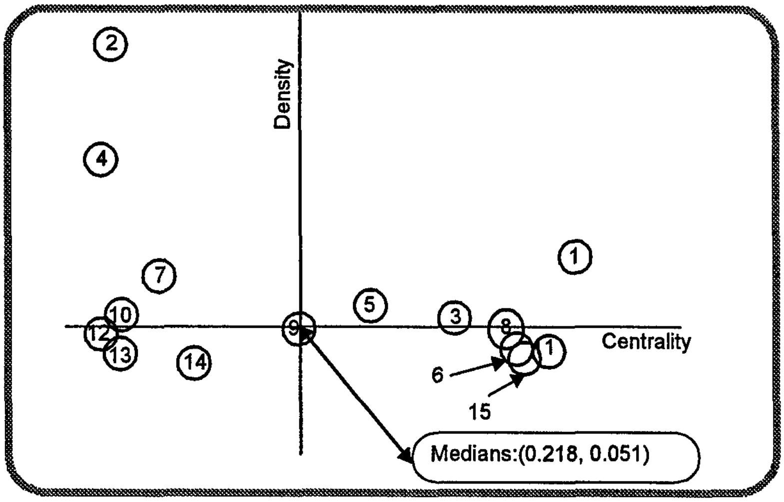


FIG. 2. 1982–1986 centrality and density.

Isolated networks show relatively low link and node counts (see 1991–1994 Network-2, Petri Nets, for example).

Perhaps the most interesting networks are the ones with both strong density and strong centrality. Few of these emerge, which testifies to software engineering's somewhat indefinite focus. None are identified in 1987–1990. However, in the 1991–1994 data, Networks-1, -3, and -4 have these properties. These networks also have strong interaction with each other. Network-7 shows strong centrality. Network-2 shows strong density but weak (actually, zero) centrality values; note that it is a completely isolated network, so its position matches intuition. Network-10 and Network-11 are below the median for both centrality and density scores. Similar analyses can be performed on the other periods.

### Super Network Analysis

In addition to describing how networks compare within a period, we can be more specific in describing how networks interact with other specific networks; this addresses centrality in a more focused fashion, but does not substitute for the general centrality measure. Centrality is a composite measure of a network's intersection with all other networks generated from the same descriptor data; super networks describe pair-wise links of networks from those data.

More formally, we defined principal and secondary networks as follows: If Network-A has internal nodes that are Pass-2 nodes in $x$ links of Network-B, and each of these $x$ links has a Pass-2 $S$ value that exceeds the minimum Pass-1 $S$ value of Network-B, then Network-A is a secondary network of Network-B. Super network (directed) links to capture this situation are of the form

$$x\text{: Network-A} \rightarrow \text{Network-B.}$$

Notice that Network-A and Network-B can share more than $x$ nodes.

Using this way of determining principal and secondary networks, we can describe super networks of networks. The

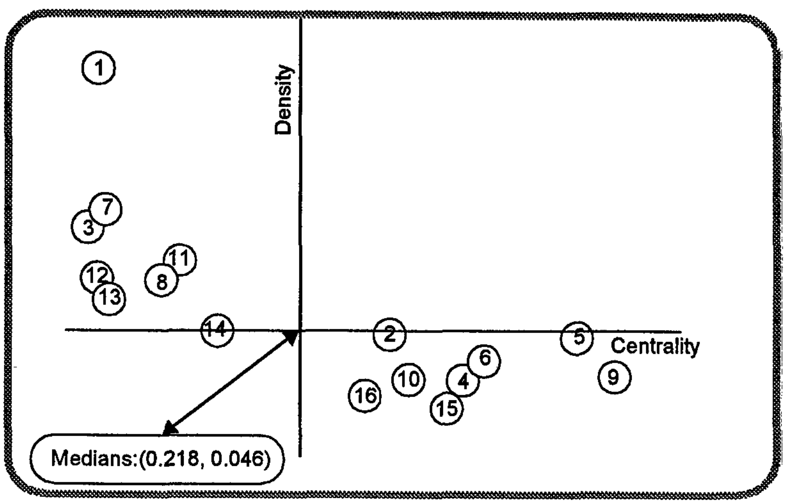


FIG. 3. 1987–1990 centrality and density.

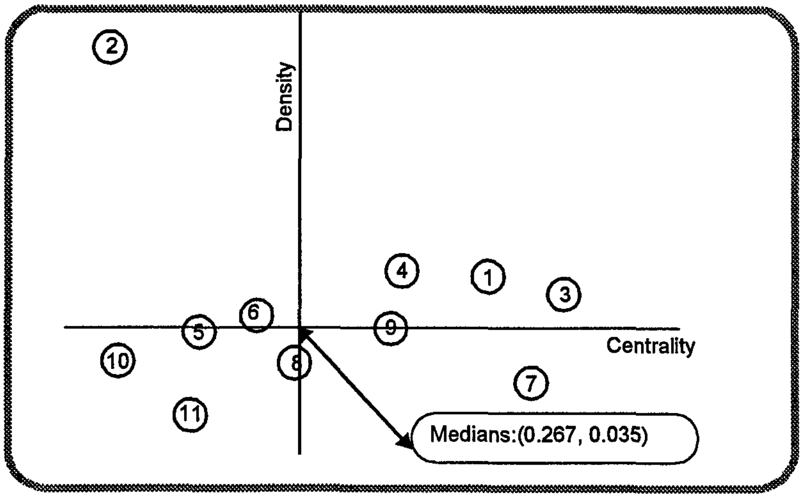


FIG. 4. 1991–1994 centrality and density.

relationships in these super networks are not inherently bi-directional, as are network links (at least as defined using $S$).

*1991–1994 Super Networks*

Table 5 gives all 1991–1994 networks that have at least one qualifying connection with other networks. Shown with each network is an entry in the form $y(z)$; $y$ indicates the associated network and $z$ shows the number of qualifying links. From this, we can then construct a super network at whatever threshold of $x$ we choose.

Setting the threshold at $x = 2$ qualifying connections, we can construct a super network of networks for each period. Consider the 1991–1994 super network (Fig. 5) and its underlying generating data (Table 5). (Fig. 5 has no scale or axes.)

Some observations include:

1. Network-2, -5, -6, -10, and -11 are isolated networks.
2. Network-3 is a secondary network of principal network Network-8; Network-7 is a secondary network of Network-8.
3. Network-3 is a principal network and a secondary network relative to both Network-4 and Network-7.
4. Network-7 is especially strongly connected to Network-3; Network-3 is less strongly connected to Network-7, at least relative to the former.

TABLE 5. Possible 1991–1994 super networks.

| Network | Connected networks [network number (number of links)] |
|---|---|
| 1 | 7 (1), 9 (2) |
| 2 | none |
| 3 | 4 (2), 7 (5) |
| 4 | 3 (3), 6 (1) |
| 5 | none |
| 6 | 4 (1) |
| 7 | 1 (1), 3 (13) |
| 8 | 1 (1), 3 (5), 7 (3) |
| 9 | 1 (2) |
| 10 | none |
| 11 | none |

Putting this in context of the networks' contents, we might conclude that:

1. Object-oriented programming is a major focus of software development.
2. Software libraries have combined with object-oriented methodologies as principal development activities.
3. The major systems used now in software engineering are Ada, C++, C, Unix, X-windows, and Microsoft Windows.
4. Computer-aided software engineering and object-oriented languages are emerging as specific tools in software development.

Looking further at the isolated networks and the Centrality/Density diagram, we might conclude that Petri Nets is either an emerging or dying research topic because it is completely isolated from other networks. Many other conclusions and impressions are derivable from the networks and super networks.

## Descriptor Contexts in Different Time Periods

Through use of network names, we observe that software development (which includes management), user interfaces, parallelism, verification and validation, requirements/specifications, and tools and techniques to be foci of study in each period. However, while these foci maintain some of the same connections over different time periods, they also evolve by forming new connections to different nodes. For

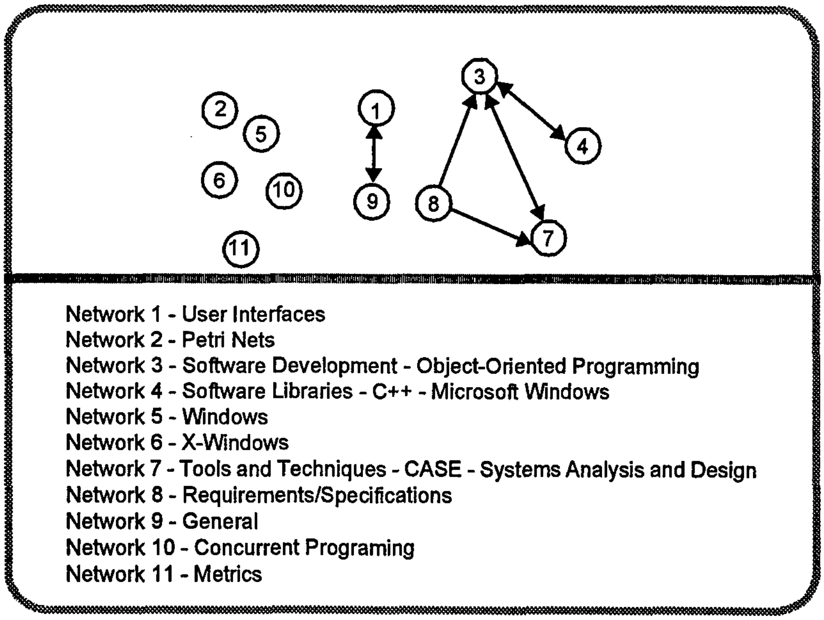


FIG. 5. 1991–1994 super network, $x = 2$.

example, in the 1991–1994 Network-7, Tools and Techniques, appears with CASE, objected-oriented techniques, reuse, and Ada, whereas in the related 1982–1986 Network-5, Tools and Techniques appears with Pascal and Structured Programming topics.

Much of the change can be gleaned from detailed examinations of networks. To illustrate this process, we will present two detailed cases.

*Software Development's Evolution*

The descriptor Software Development (K.6.3) forms at least part of the titles of network in each period of study. It is a major theme in 1982–1987 Network 6—Software Development, 1987–1990 Network 6—Methodologies—Software Development, and 1991–1994 Network 3—Software Development—Object Oriented Programming, respectively. However, the context of the node changes over time periods, as Software Development K.6.3 links evolve from ones with many non-specific, encompassing descriptors to ones with more specific relationships to tools, methodologies, processes, and people concerns. At the same time, Software Development K.6.3 consistently associates with some fundamental concepts, such as Management D.2.9. Table 6 displays these contexts over time. Descriptors in bold print appear for the first time in association with

TABLE 6. Links with software development K.6.3.

| 1982–1986 | 1987–1990 | 1991–1994 |
|---|---|---|
| **Design D.2.10** | **Tools and techniques D.2.2** | Design D.2.10 |
| **Programming environments D.2.6** | Programming environments D.2.6 | Programming environments D.2.6 |
| **Systems development K.6.1** | Systems development K.6.1 | **Object oriented programming D.1.5** |
| **General D.2.0** | **Methodologies D.2.10** | Methodologies D.2.10 |
| **Life cycle D.2.9** | Life cycle D.2.9 | **Computer-aided software engineering (CASE) D.2.2** |
| **Management D.2.9** | Management D.2.9 | Management D.2.9 |
| **Testing and debugging D.2.5** | **Software configuration management D.2.9** | Software configuration management D.2.9 |
| **Systems analysis and design K.6.1** | **Documentation D.2.7** | Systems analysis and design K.6.1 |
| **General D.1.0** | **Documentation D.2.7** | **Software quality assurance (SQA) D.2.9** |
| | | **Programming teams D.2.9** |
| | | **Human factors H.1.2** |

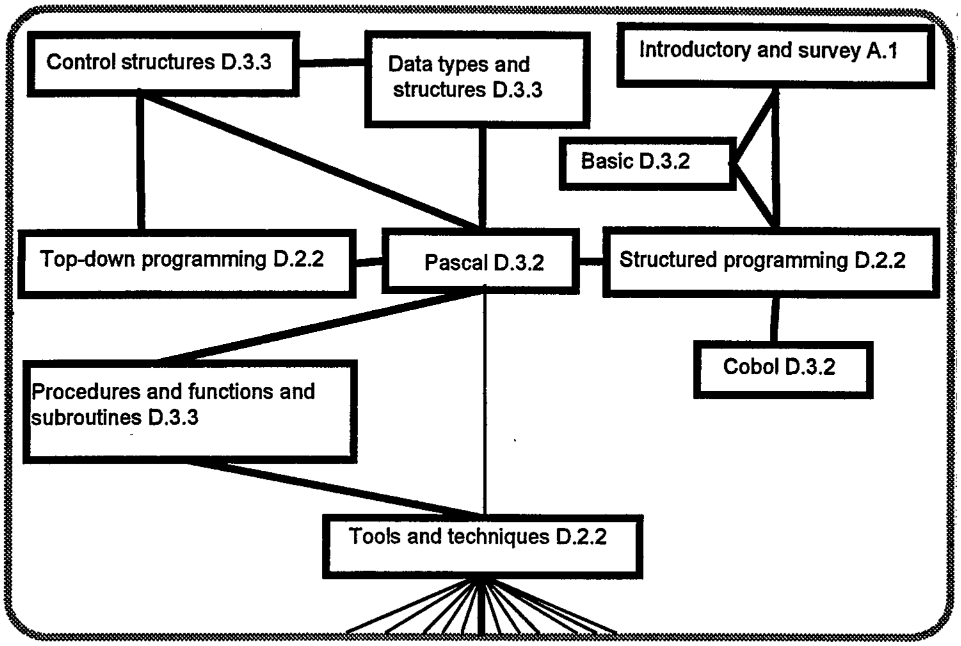


FIG. 6. Structured programming in Network-5, 1982–1986.

Software Development K.6.3. Recurring descriptors are aligned horizontally when possible, although a recurring descriptor may be absent in one time period; otherwise, the order of entry is arbitrary.

*Structured Programming's Evolution*

We demonstrate structured programming's transformation through the two networks named after descriptor Structural programming D.2.2, 1982–1986 Network 5—Tools and Techniques—Structured Programming—Pascal and 1987–1990 Network 8—Structured Programming. Figures 6 and 7 trace this evolution through full depictions of the networks.

As high-level software issues become more integrated, older ones fade. Pascal, Basic, and Cobol appear in 1982–1986 Network-5 (Tools and Techniques—Structured Programming—Pascal). This network is based on programming-in-the-small issues, such as structured programming and top-down programming. In 1987–1990 Network-8 (Structured Programming), Basic and Cobol appear almost in isolation with structured programming. That theme then disappears in 1991–1994 as software engineering research moves to programing-in-the-large concerns.

Other similar context analyses are possible. For example, the GUIDE data reveal the evolution of the programming language Ada (Coulter, Monarch, Konda, & Carr, 1995).

## Similarity of Networks in Different Time Periods

The transformation of networks and their intersections with other networks across time periods provides insights into the emergence of software engineering research

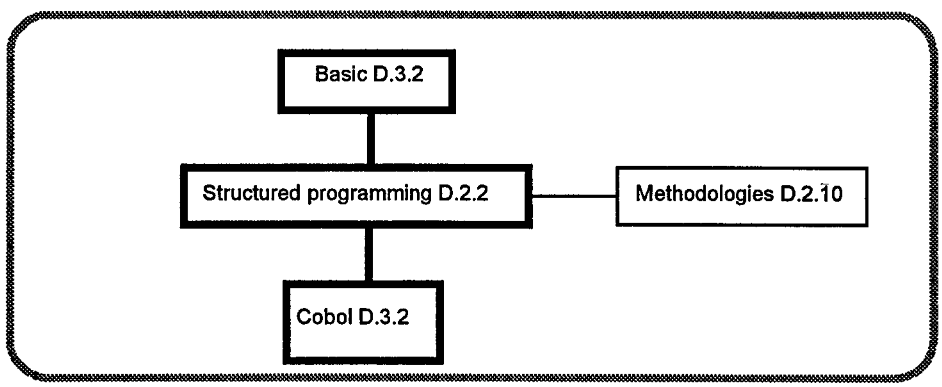


FIG. 7. Structured programming in Network-8, 1987–1990.

themes. To quantify this analysis, we apply the *Similarity Index* (*SI*) approach, which is patterned after Callon's Dissimilarity Index (Callon et al., 1991).

*SI* measures the intersection of the descriptors in two networks. It does not directly include the corresponding links in networks; however, since all descriptors in a network are at least indirectly liked, this metric captures some portion of network similarity.

Consider two networks $N_i$ and $N_j$. Let $w_i$ be the number of descriptors in $N_i$, let $w_j$ be the number of descriptors in $N_j$, and let $w_{ij}$ be the number of descriptors common to $N_i$ and $N_j$. Then,

$$SI(w_i, w_j, w_{ij}) = 2 \times \left(\frac{w_{ij}}{w_i + w_j}\right),\ 0 < SI \leq 1.$$

We multiply by 2 so that the maximum value of *SI* is 1, which occurs when $N_i$ and $N_j$ have identical nodes.

### *Emergence of 1991–1994 Themes*

We can apply *SI* to examine the emergence of some 1991–1994 networks. Especially interesting are the three networks showing both strong centrality and density values (called core networks or core themes). The networks are Network-1, User Interfaces; Network-3, Software Development—Object-Oriented Programming; and Network-4, Software Libraries—C++—Microsoft Windows.

First, consider 1991–1994 Network-1, User Interfaces. It has reportable *SI* intersections with four 1987–1990 networks, as shown below.[10]

| $N_1$ | $N_2$ | $w_1$ | $w_2$ | $w_{12}$ | *SI* |
|---|---|---|---|---|---|
| **1991-1994 Network 1** | **1987-1990 Networks** | | | | |
| User Interfaces | 1. Network-5: User/Machine Systems | 14 | 14 | 6 | 0.423 |
| | 2. Network-6: Methodologies - Software Development | 14 | 20 | 7 | 0.412 |
| | 3. Network-9: Applications and Expert Systems - Tools and Techniques | 14 | 19 | 5 | 0.303 |
| | 4. Network-16: Software Management | 14 | 11 | 5 | 0.400 |

Hence, the 1991–1994 theme User Interfaces incorporates descriptors from several 1987–1990 networks. Its emergence history is complicated; tracing it further could require investigation of four 1987–1990 networks and of all their 1982–1986 predecessor networks.

Similarly, 1991–1994 Network-3, Software Development—Object-oriented Programming, displays a multiply engendered network history. It has reportable *SI* values with seven 1987–1990 networks.

| $N_1$ | $N_2$ | $w_1$ | $w_2$ | $w_{12}$ | *SI* |
|---|---|---|---|---|---|
| **1991-1994 Network 3** | **1987-1990 Networks** | | | | |
| Software Development - Object-Oriented Programming | 1. Network-4: Requirements/ Specification - Methodologies | 18 | 16 | 5 | 0.294 |
| | 2. Network-6: Methodologies - Software Development | 18 | 20 | 7 | 0.369 |
| | 3. Network-9: Applications and Expert Systems - Tools and Techniques | 18 | 19 | 5 | 0.270 |
| | 4. Network-10: Concurrent Programming - Ada | 18 | 18 | 7 | 0.389 |
| | 5. Network 14: Software Configuration Management | 18 | 6 | 6 | 0.500 |
| | 6. Network 15: Reusable Software | 18 | 16 | 7 | 0.417 |
| | 7. Network 16: Software Management - Design | 18 | 11 | 8 | 0.552 |

Observe that 1987–1990 Network-14, Software Configuration Management, was completely absorbed by the 1991–1994 network (i.e., all descriptors of the earlier network are descriptors of the latter network).

Now, consider 1991–1994 Network-4 Software Libraries—C++—Microsoft Windows. It has a reportable *SI* value, 0.375, for only one 1987–1990 network, Network-15, Reusable Software. Tracing this latter network to 1982–1986 ones shows that it has a reportable *SI* value, 0.343, only for 1982–1986 Network-6, Software Development. This 1987–1990 network also absorbs 1982–1986 Network-7, Software Libraries, and Network-12, Modules and Interfaces; but, each of these networks has fewer than five descriptors, so criteria for *SI* scores are not met. This history suggests a relatively well-defined emergence path for themes dealing with software reuse.

*SI* analysis can also show the lack of a traceable past. Consider 1991–1994 Network-6, X-Windows. It has no identifiable 1987–1990 predecessors. Only four networks from that earlier period share even one descriptor with it (in all cases the same descriptor—(User interfaces D.2.2)). Similarly, 1991–1994 Network-5, Windows, has no reportable 1987—1990 predecessors. Only two 1987–1990 networks share any descriptors with it (1987–1990 Network-10 and -15, with one and two descriptors, respectively). Taken together, we see a rapid emergence of windows-based research. Sometimes research foci emerge quickly, as expected in a dynamic field.

*Similarity Index within a Time Period*

*SI* can also be useful within a time period to assess the similarity of companion networks. Consider the 1991–1994 core networks; they have substantial intersection with each other, as seen below:

| $N_1$ | $N_2$ | $w_1$ | $w_2$ | $w_{12}$ | *SI* |
|---|---|---|---|---|---|
| **1991-1994 Network 1** | **1987-1990 Networks** | | | | |
| User Interfaces | 1. Network-5: User/Machine Systems | 14 | 14 | 6 | 0.423 |
| | 2. Network-6: Methodologies - Software Development | 14 | 20 | 7 | 0.412 |
| | 3. Network-9: Applications and Expert Systems - Tools and Techniques | 14 | 19 | 5 | 0.303 |
| | 4. Network-16: Software Management | 14 | 11 | 5 | 0.400 |

The network predecessors of 1991–1994 core themes demonstrate notable characteristics. All of them with reportable *SI* scores also have high centrality scores (Figs. 2, 3, and 4) for the time periods of interest, except for 1987–1990 Network-14, which had a slightly below median score. However, that network was completely absorbed by its successor. Similarly, two 1982–1986 networks with below median centrality scores were completely absorbed by their successor, even though their *SI* scores were not reportable. In these latter cases, the networks were all small and relatively isolated.

This observation suggests that core themes may normally emerge from predecessor networks that already display relatively strong connections to other networks within the same time period. It also suggests that isolated networks may quickly become part of more integrated networks in a succeeding time period. This absorption could occur because one new link connects a small, isolated network to a larger network. However, certainly not all isolated networks merge with larger ones, as is so far evident of the Standards networks of 1982–1986 and 1987–1990. As noted above in the case of the Structural Programming theme, a network also can transform from a core theme (1982–1986, Network-5) to an isolated theme (1987–1990, Network-8).

## Descriptor Analysis

Analysis of descriptor nodes gives a supporting view of which descriptors in CCS—but outside of Software Engineering—interact with software engineering descriptors. Recall that only descriptors which co-occur with other descriptors a requisite number of times, and with relatively high strength, are candidates for inclusion in networks. Many descriptors that appear in documents do not associate often enough or strongly enough with other descriptors to be considered for inclusion. The strengths of associations relative to other associations further limits which links enter into a network. Of the 1,606 unique descriptors appearing in all documents, 158 (9.8%) de-

[10] Only CCS descriptors defined in both pertinent time periods are included in *SI* descriptor counts. To insure notable intersection between $N_i$ and $N_j$, we require $w_{ij} \leq 5$ before reporting *SI*.

TABLE 7. Summary of descriptor data.*

| | Rank order of descriptor statistics by generation | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Rank in no. documents | | | Rank in no. networks | | | Rank in no. times in networks | | |
| Descriptor | 82–86 | 87–90 | 91–94 | 82–86 | 87–90 | 91–94 | 82–86 | 87–90 | 91–94 |
| Ada D.3.2 | 8 | 13 | 15 | 7t | 10t | 3t | 13 | 6t | 7 |
| Applications and expert sys I.2.1 | — | 15 | — | — | 6t | — | — | 5 | — |
| Computer aided . . . (case) D.2.2 | # | # | 11 | # | # | 11t | # | # | 8t |
| Correctness proofs D.2.4 | 14 | — | — | 11t | — | — | 10t | — | — |
| Design D.2.1 | # | 14 | — | # | 10t | — | # | 14 | — |
| General D.2.0 | 3 | 7 | 7 | 7t | 10t | 11 | 10t | 12t | 15 |
| Human factors H.1.2 | 6 | 8 | — | 7t | 6t | — | 10t | 12t | — |
| Interactive D.2.6 | — | 12 | — | — | 10t | — | — | 6t | — |
| Management D.2.9 | 9t | — | 13 | 4t | — | 5t | 5t | — | 8t |
| Methodologies D.2.10 | # | 3 | 10 | # | 1 | 11t | # | 1 | 12t |
| Metrics D.2.8 | 15 | — | — | 7t | — | — | 8t | — | — |
| Object-oriented programming D.1.5 | # | # | 2 | # | # | 2 | # | # | 1t |
| Program verification D.2.4 | 13 | — | — | 11t | — | — | 14t | — | — |
| Programming environments D.2.6 | 2 | 4 | 6 | 1t | 4 | 8t | 5t | 9t | 12t |
| Requirements/specifications D.2.1 | 12 | 6 | 8 | 6 | 6t | 15 | 4 | 6t | 6 |
| Reusable software D.2.m | — | 9 | 9 | — | 6t | 5t | — | 4 | 8t |
| Software development K.6.3 | 4 | 5 | 5 | 1t | 2 | 3t | 1 | 3 | 1t |
| Software management K.6.3 | 9t | — | — | 1t | — | — | 2t | — | — |
| Structured programming D.2.2 | 11 | — | — | 15 | — | — | 14t | — | — |
| Testing and debugging D.2.5 | 7 | 10 | 12 | 4t | 15 | 11t | 8t | 15 | 8t |
| Tools and techniques D.2.2 | 5 | 2 | 3 | 7t | 3 | 1 | 2t | 2 | 4 |
| User interfaces D.2.2 | 1 | 1 | 4 | 11t | 5 | 8t | 5t | 9t | 5 |
| User interfaces H.5.2 | # | # | 14 | # | # | 8t | # | # | 12t |
| User/machine systems H.1.2 | — | 11 | 2 | — | 10t | — | — | 9t | — |
| Windows D.2.2 | — | — | 1 | — | — | 5t | — | — | 1t |

*# = Node not in CCS for period. t = Tie for ranked position. Ties for position $n$ all ranked as $n$; next ranked position begins at $n + m$, where $m$ is number of ties ranked at $n$. — = Not in highest 15 for period.

scriptors satisfied these criteria and appeared in the generated networks. We cannot define the maximum possible number of nodes because of unrestricted numbers of implicit subject descriptors.

Table 7 summarizes the most frequently appearing descriptors in each time period. The table was generated by first obtaining the 15 most frequently appearing descriptors within each time period and then eliminating redundancy from the combine lists. The descriptors are listed alphabetically in Table 7. For each descriptor, its rank in each period is shown by the number of documents in which it appears, the number of networks in which it appears, and the number of times it appears (a descriptor can be connected to more than one other descriptor in the same network, as evident in Fig. 1 for the descriptor Management D.2.9).

Some common themes also emerge from descriptor data. Tools and techniques, user interfaces, programming environments, reusable software, design methodologies, software management and development, testing and debugging, verification, metrics, Ada, and requirements/ specifications appear consistently and repeatedly. Some new descriptors are prominent in 1991–1994 data, including computer-aided software engineering, object-oriented programming, and Windows.

Only the following 25 descriptors appeared in all time periods (not just among the 15 most common by period).[11]

Ada D.3.2
Concurrent programming D.1.3
Curriculum K.3.2
Design D.2.10
General D.2.0
Human factors H.1.2
Interaction techniques I.3.6

[11] Recall that new CCS descriptors created in 1987 and 1991 are not candidates for appearance in preceding time periods. Hence, some now commonly used descriptors—such as object-Oriented Programming D.1.5—could not appear in this list.

TABLE 8. CCS descriptor summary data.*

| CCS category | All descriptors (57,725) (%) | Unique descriptors (1,606) (%) | Network descriptors (158) (%) |
|---|---|---|---|
| A—General literature | 0.6 | 0.4 | 1.3 |
| B—Hardware | 1.1 | 7.5 | 0 |
| C—Computer systems organization | 4.4 | 8.5 | 3.2 |
| D—Software | 59.1 (40.1, in D.2) | 31.9 (11.2 in D.2) | 58.3 (29.1 in D.2) |
| E—Data | 0.6 | 1.8 | 0 |
| Theory of computation | 4.5 | 5.2 | 6.3 |
| G—Mathematics of computing | 1.7 | 5.3 | 1.9 |
| H—Information systems | 8.9 | 11.6 | 8.9 |
| I—Computing methodologies | 7.6 | 16.1 | 12.0 |
| J—Computer applications | 2.5 | 3.6 | 1.3 |
| K—Computing milieux | 8.9 | 8.1 | 11.3 |

* The B and E CCS categories were not represented at all in the networks, and the A, C, G, and J categories were only marginally included. The F category was included primarily with respect to program verification.

Introductory and survey A.1
Management D.2.9
Mathematical software G.4
Methodologies D.2.1
Metrics D.2.8
Program verification D.2.4
Programming environments D.2.6
Requirements/specifications D.2.1
Software development K.6.3
Software libraries D.2.2
Software management K.6.3
Software quality assurance (sqa) D.2.9
Specification techniques F.3.1
Specifying and verifying and reasoning about programs F.3.1
Testing and debugging D.2.5
Tools and techniques D.2.2
User interfaces D.2.2
User/machine systems H.1.2

Another way to see the filtering effect of the algorithm is to count the descriptors in each major CCS category in the original data and compare that number to the ones that emerged as network nodes.

Table 8 gives the percentage of the 57,727 descriptors in the original data by CCS category (first column), the percentages of the 1,606 unique descriptors in the original data by CCS category (second column), and the percentages by CCS category of the 158 descriptors that passed the co-word analysis filter to reach the resulting 42 networks.

Listed below are the eight non-D.2 descriptors included among the 15 most frequent descriptors in networks (Table 7). These descriptors highlight interactions among D.2 descriptors and other descriptors in CCS. As evident there, much of software engineering's intersection with the rest of computing is in the areas of user interaction, software management, and programming methodology. Very little interaction with hardware, data, mathematics of computing, and computer applications is evident. Further analyses will reinforce this hypothesis.

D.1.5—Object-oriented programming
D.3.2—Ada
H.1.2—Human factors
H.1.2—User/machine systems
H.5.2—User interfaces
I.2.1—Applications and expert systems
K.6.3—Software development
K.6.3—Software management

Our analysis of the CCS descriptors shows that some D.2 descriptors play a less important role than expected and implied in some software engineering definitions (Denning et al., 1989; Humphrey, 1989; IEEE, 1989; Shaw, 1990). These definitions normally incorporate terms such as large scale, economical, managerial, interdisciplinary, production, maintenance, reliable, dependable, efficient, safety, design, and specifications. We see some of these themes in our findings, but not all of them.

Human factors is a consistent and important theme in all periods we analyzed; that part of software engineering seems marginal in other attempts to define it. Conversely, economic aspects are mentioned consistently in these other discussions; yet we found little on that subject in the research and development literature, even though descriptors under (D.2.9) MANAGEMENT—Cost Estimation, and—Time Estimation would accommodate the matter—as would (K.6.0) GENERAL—Economics. Over the three time periods analyzed here, these descriptors appeared in the unfiltered data 117 times, 15 times, and 33 times, respectively, but did not associate strongly enough with other descriptors to be placed in any networks.

Also, we find little evidence of a maturing profession as judged by commentary on issues such as ethics, licensing, certification, human safety, and codes of good practice, even though appropriate CCS nodes are defined. None of these nodes reached the networks, and only minimal inclusion was found in the almost 58,000 total, unfiltered descriptors. While the standards descriptors were includes in the first two generations of networks (but in isolated fashions), they

did not appear in 1991–1994 networks. As stated by Shaw (1990), an engineering discipline of software is yet in early stages of development.

## Conclusions

What can we conclude about the state of software engineering based on our study of publications? First, the field is rapidly evolving, as is demonstrated by the changing descriptors in networks and by the changing centrality/density scores. The analysis of the 1991–1994 data shows a trend towards focusing on object-oriented themes, software reuse/software library themes, and user interface themes. Consistent themes are evident over the time periods studied, although contexts change. Some consistent themes are user interfaces, tools and techniques, verification and validation, software reuse, requirements and specifications, and design methodologies.

While some software engineering themes seem to have well-defined genealogies, others appear to emanate from multiple preceding themes; still others emerge quickly with little evidence of ancestry. So, the field has some established research themes, but it also changes rapidly to embrace new themes. The core themes of user interfaces and software development (with object-oriented methods) both display underlying principles (such as screen design, design methodologies, reusable software, and so forth) together with software tools that embody some of these underlying principles. These tools include X-Windows, Microsoft Windows, Ada, C++, and UNIX. CASE tools are prominent in software development networks, but names of specific CASE tools are not present. This observation suggests that the maturity of a software engineering subfield can be gauged by the maturity of relevant supporting tools. Earlier, we observed that the languages Pascal, Basic, and Cobol dropped from the software engineering descriptors, along with programming-in-the-small issues such as structured programming. They were replaced by programming-in-the-large issues and by a different set of supporting tools appropriate for large-scale software development environments. As software engineering matures, we can expect to observe the names of other specific software tools and systems, and we may see new core areas emerge as supporting tools are refined.

Because CCS is a fixed taxonomy with periodic updates to descriptors, the role of implicit subject descriptors may be crucial in observing trends between and across updates to the classification system. Therefore, names of languages and systems as reflected in CCS descriptors provide numerous insights into observing a field's maturation.

This study demonstrates the feasibility of co-word analysis as a viable approach for extracting patterns from, and identifying trends, in large corpora where the texts collected are from the same subdomain and divided into roughly equivalent quantities for different time periods. Hence, this examination, of software engineering's emergence, points the way for further research into the evolution of software engineering and computing, as well as well-restricted subdomains in other technical fields.

## Acknowledgments

Marvin Carr, David Gluch, Allen Tucker, Nancy Mead, Robert Park, Will Hayes, Eswaran Subrahmanian, Mario Barbacci, Anthony Ralston, Christopher Fox, and Dennis Frailey provided many helpful suggestions for improving the content and clarity of this article. Carol Wierzbicki, Craig Rodkin, and Bernard Rous of ACM provided access to data for this research. We thank the anonymous referees for their helpful comments.

## References

Boehm, B. (1994). The IEEE-ACM initiative on software engineering as a profession. *Software Engineering Technical Council Newsletter, 13(1),* 1.

Buckley, F. (1993). Defining software engineering. *Computer, 26(8),* 76–78.

CR 1996. (1996). Periodicals received. *Computing Reviews, 36(11),* 599–608.

CR 1997. (1997). The full computing reviews classification system. *Computing Reviews, 37(1),* 4–16.

Callon, M., Courtial, J.-P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics, 22(1),* 153–203.

Callon, M., Law, J., & Rip. A. (1986). *Mapping of the dynamics of science and technology.* London: Macmillan.

Chen, H. (1994). Collaborative systems: Solving the vocabulary problem *IEEE Computer, 27(5),* 58–66.

Coulter, N. (1998). Changes to the computing classification system. *Computing Reviews, 39(1).*

Coulter, N., & Dammann, J. (1994). Current practices, culture changes, and software engineering education. *Computer Science Education, 5,* 91–106.

Coulter, N., & Monarch, I. (1996). Best practices: What software engineering can learn from manufacturing engineering, *Proceedings of the 1996 SEI Software Engineering Symposium: Achieving Maturity through Technology Adoption* (pp. 7–8) Pittsburgh, September 9–12: Pittsburgh: SEI.

Coulter, N., Monarch, I., Konda, S., & Carr, M. (1995). Ada and the evolution of software engineering. In G. Engle (Ed.), *Proceedings of TRI-Ada'95: Solutions for a Changing, Complex World* (pp. 56–71) Anaheim, CA: ACM.

Coulter, N., Monarch, I., Konda, S., & Carr, M (1996). *An evolutionary perspective of software engineering research through co-word analysis* (Tech. Rep. No. CMU/SEI-96-TR-019). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University.

Courtial, J.-P. (1994). A co-word analysis of scientometrics. *Scientometrics, 31(3),* 251–260.

Courtial, J.-P., Cahlik, T., & Callon, M. (1994). A model for social interaction between cognition and action through a key-word simulation of knowledge growth. *Scientometrics, 31(2),* 173–192.

Courtial, J.-P., & Law, J. (1989). A co-word study of artificial intelligence. *Social Studies in Science, 19,* 301–311.

Denning, P., Gries, D., Mulder, M., Tucker, A., Turner, J., & Young, P. (1989). Computing as a discipline. *Communications of the ACM, 32(1),* 9–23.

Ford, G., & Gibbs, N. (1989). A master of software engineering curriculum. *Computer, 22(9),* 59–71.

Gaines, B. (1993). An agenda for digital journals: The socio-technical infrastructure of knowledge dissemination. *Journal of Organizing Computing, 3(2),* 135–193.

Gaines, B. & Shaw, M. (1995). Knowledge acquisition and representation techniques in scholarly communication. *Proceedings of SIGDOC '95 Conference. Emerging from Chaos: Solutions for the Growing Complexity of Our Jobs. (13th Annual International Conference on Systems Documentation Conference,* Savannah, GA, October 2–4). 197–206, New York: ACM.

Garfield, E. (1983). *citation indexing: Its theory and application in science, technology, and humanities* (2nd ed.). Philadelphia: ISI Press.

Gibbs, N. (1989). The SEI education program: The challenge of teaching future software engineers. *Communications of the ACM, 32(5),* 594–605.

Gibbs, N. (1991). Software engineering and computer science: The impending split. *Education and Computing, 7,* 111–117.

Humphrey, W. (1989). *Managing the software process.* Reading, MA: Addison-Wesley.

IEEE (1989). *Software engineering standards.* New York: IEEE.

Kessler, M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14,* 10–25.

Law, J., & Whittaker, J. (1992). Mapping acidification research: A test of the co-word method. *Scientometrics, 23(3),* 417–461.

Levy, D. (1995). Cataloging in the digital order. *DL'95: The Second Annual Conference on the Theory and Practice of Digital Libraries,* Austin, TX. Available at: http://csdl.tamu.edu/DL95/contents.html.

Liu, M. (1993). The complexities of citation practice: A review of citation studies. *Journal of Documentation, 49,* 370–408.

Monarch, I. (1994). An interactive computational approach for building a software risk taxonomy. *Proceedings of The Third SEI Conference on Software Risk,* (J. Allen, Ed.) April 5–7, Pittsburgh, PA. Software Engineering Institute: Pittsburgh, PA.

Monarch, I. (1996). Software engineering risk repository. *Proceedings of 1996 SEPG Conference,* (J. Allen, Ed.) May 20–23, Atlantic City, New Jersey. Software Engineering Institute: Pittsburgh, PA.

Monarch, I., & Gluch, D. (1995). *An experiment in software development risk analysis* (Software Engineering Tech. Rep. No. CMU/SEI-95-TR-014). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University.

Monarch, I., Konda, S., Levy, S., Reich, Y., Subrahmanian, E., & Ulrich, C. (1997). Mapping sociotechnical networks in the making. In J. Bower, L. Gasser, L. Star, & W. turner (Eds.), *Social science, technical systems, and interactive technology* (pp. 331–354). Mahwah, NJ: Lawrence.

Monarch, I., & Nirenburg, S. (1993). An ontological approach to knowledge acquisition schemes. In N. Bourbakis (Ed). *Knowledge engineering shells: Systems and techniques* (pp. 29–56). NJ: World Scientific.

Mylopoulos, J., Borgida, A., & Yu, E. (1997). Representing software engineering knowledge. *Automated Software Engineering, 4,* 291–317.

Parnas, D. (1990). Education for computer professionals. *Computer, 23(1),* 17–22.

Parnas, D. (1997). Software engineering: An unconsummated marriage. *Communications of the ACM, 40(9),* 128.

Savoy, J. (1997). Ranking schemes in hybrid Boolean systems: A new approach. *Journal of the American Society for Information Science, 48,* 235–253.

Schatz, B., & Chen, H. (1996). Building large-scale digital libraries, IEEE Computer, *29*(5), 22–26.

Schatz, B., Johnson, E., Cochrane, P., & Chen, H. (1996a). Interactive term suggestion for users of digital libraries: Using subject thesauri and co-occurrence lists for information retrieval. In E. Fox & G. Marchionini (Eds.), *Proceedings of the First ACM International Digital Libraries Conference* (pp. 126–133). New York: ACM Press.

Shatz, B., Mischo, W., Cole, T., Hardin, J., Bishop, A., & Chen, H. (1996b). Federating diverse collections of scientific literature. *IEEE Computer, 29(5),* 28–36.

Shaw, M. (1990). Prospects for an engineering discipline of software. *Software, 6(6),* 15–24.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between documents. *Journal of the American Society for Information Science, 24,* 265–269.

Small, H., & Griffith, B. (1974). The structure of scientific literature I: Identifying and graphing specialities. *Science Studies, 4,* 17–40.

Tucker, A. (1991). Computing curricula 1991. *communications of the ACM, 34(6),* 68–84.

Turner, W., Chartron, G., Laville, F., & Michelet, B. (1988). Packaging information for peer review: New co-word analysis techniques. In A. Van Raan (Ed.), *Handbook of quantitative studies of science and technology* (pp. 291–323). Amsterdam: North Holland.

Turner, W., Lelu, A., & Georgel, A. (1994). Geode: Optimizing data flow representation techniques in a network information system. *Scientometrics, 30(1),* 269–281.

Turner, W., & Rojouan, F. (1991). Evaluating input/output relationships in a regional research network using co-word analysis. *Scientometrics, 22(1),* 139–154.

Weinberg, B. (1974). Bibliographic coupling: A review. *Information Storage and Retrieval, 10,* 189–196.

Whittaker, J. (1989). Creativity and conformity in science: Titles, keywords, and co-word analysis. *Social Science in Science, 19,* 473–496.

Zitt, M. (1991). A simple method for dynamic scientometrics using lexical analysis. *Scientometrics, 22(1),* 229–252.

Zitt, M., & Bassecoulard, E. (1994). Development of a method for detection and trend analysis of research fronts built by lexical and cocitation analysis. *Scientometrics, 30(1),* 333–351.

# Information Aspects of New Organizational Designs: Exploring the Non-Traditional Organization

**Bob Travica**
*School of Library and Information Science, Indiana University, 10th & Jordan, Bloomington, IN 47405. E-mail: btravica@indiana.edu*

**The purpose of the study presented in this article aims at broadening our understanding of information and concomitant aspects of a non-bureaucratic organizational design. With current changes in organizational environments, the century-long domination of the bureaucratic organization is being shaken. New organizational designs have been proposed as alternatives to the bureaucracy, including the information-based organization, networked organization, and adhocracy. Our knowledge of these designs, however, is still meager. This particularly applies to their information aspects, such as the role of information technology (IT), and exchanges of information and knowledge. In order to fill the void, an organizational design which will be called the "non-traditional organization" was created on the basis of relevant literature and was preliminarily tested on a sample drawn from the public accounting industry. The study discovered a strong positive relationship between the amount of IT usage and non-traditional dimensions, invoking the notion of an organic, "informated" organization. This relationship primarily rests on the relationships IT forms with centralization and formalization (both negative), and trust and communication beyond team boundaries (both positive). Information and knowledge-related interactions are in relation with team-based accountability in this networked, "information-rich" organization. Communication beyond team boundaries serves several purposes, including integration and creation of a dialogue-based, adhocratic, "interactive" organization. Hierarchy appears to be the most resilient dimension of the traditional organization in the sample studied. These findings bear implications for understanding new organizational designs.**

## 1. Research Problem

The century-long domination of the bureaucratic organization in U.S. business is being shaken. From Drucker's (1987/1990) laconic acknowledgment of this thesis, to its more recent comprehensive treatments (e.g., Heckscher & Donnellon, 1994; Taylor & Van Every, 1993), a number of prominent authors have concurred that the days of the bureaucracy are numbered. Donnellon and Scully (1994) concur, arguing that the advantages of efficiency and control that the bureaucracy has historically exhibited are no longer paying off when product quality, short product time, and innovativeness are conditions for organizational survival. Not only is the bureaucracy incapable of responding to these new requirements, but it is hobbling the effort to meet them. A new organization is therefore needed, argue the authors.

The 1980s and beyond have indeed exhibited significant organizational changes. A rapid development of information technology (IT) that builds on the coupling of computing and communications capabilities is one of them. Electronic mail, for example, was born out of the marriage between the text creation capabilities of word processors and the communication capabilities of computer networks. In a similar vein, computer communications capabilities were coupled with database technology to give rise to systems for supporting group work. IT developments were paralleled by other organizational changes. In a quest for higher quality production and a more efficient organization that would capitalize on new IT, reorganization of work has been carried out (Hammer, 1996; Hammer & Champy, 1993). In addition, middle management ranks have shrunk, in part due to the reporting capabilities of IT, which resulted in flatter organizational structures (Peters, 1987; Quinn, 1992). In no small measure were organizational missions, goals, and strategies redesigned, typically emphasizing the increase in competitiveness and value for the customer (Hamel & Prahalad, 1994; Peters, 1992). New IT is sometimes brought into explicit relation with a strategic reorientation (e.g., Goldman, Nagel, & Preiss, 1995). Organizational cultures, too, were revamped. For example, the cultures that excelled in the 1970s (see Peters & Waterman, 1982) gave way to cultures that abandon the traditionally indispensable employee loyalty (Cushman & King, 1995), or to cultures extolling loosely defined work roles, and teamwork which can be supported by IT (Mankin, Cohen, & Bikson, 1996; Morgan, 1993).

Received September 2, 1996; revised September 8, 1997 and October 23, 1997; accepted November 21, 1997.

All these changes might be related with the new organization that Drucker (1987/1990) and other students of the non-bureaucratic organization talk about. Models of this organization, provided in various stages of maturation, are not in short supply. For example, Mintzberg (1979) developed the model of adhocracy, whose main characteristic is flux in work roles and other organizational aspects, as production requirements at hand demand. Myles and Snow (1986) proposed the concept of a network organization that comes into being through connecting firms specialized in particular organizational functions which exist in the typical departmentalized firm. Proposing a concept of the shamrock organization, Handy (1989) emphasizes the changing nature of the labor force, including a large segment of temporary specialist workers. Nohria and Berkley (1994) sketch the virtual organization, which can be a temporary coalition of members from different firms, dispersed in space and linked via the new IT. Quinn (1992) portrays the infinitely flat organization, with the intention to emphasize the trend of flattening organizational hierarchy and authority allocation to operational levels. Other authors propose summative properties of the new designs (Clegg, 1990; Fulk & DeSanctis, 1995; Galbraith & Lawler, 1993; Heydebrand, 1989). They, for example, agree that freer information flows and IT play an important role in these new designs, and portray structural changes in a similar way (e.g., decentralization, flattening of hierarchy, and informal networking).

New organizational designs can be viewed as descendants of the historical, non-bureaucratic blueprint—the organic form of Burns and Stalker (1961). The organic form puts professionals in charge of production, information and communication, and breaks with the rigid vertical distribution of managerial control. The common denominators of these new designs are embodied, therefore, in a negation of the bureaucratic hierarchic control, narrowly-defined work roles, segmented information, and knowledge flows and formalism, while providing alternative organizing principles, such as pushing the authority down the ladder, basing the organization of work on teams, and expanding exchanges of information and knowledge.

The coming of the new organization poses special challenges for information scientists. This organization is, first and foremost, based on information and knowledge. The technology needed for manipulating these is equally important. Moreover, not only has the management of these information resources (information, knowledge, and IT) been called into question, but due attention has to be paid to structural, cultural, political, socio-psychological, and other organizational dimensions that weave the milieu for the information resources. The relevant literature, however, is lagging behind these requirements. The empirical literature, for example, exhibits a piecemeal character—individual variables and their relationships are studied rather than organizations taken as wholes. This piecemeal evidence includes, for example, relationships between IT and structural and cultural dimensions (e.g., Carter, 1984; Kerr & Hiltz, 1982; Markus, 1994; Olson & Bly, 1991; Wijnhoven & Wassenaar, 1990), and the place and importance of information in production processes (e.g., Olson & Bly, 1991). This evidence establishes the breeding ground for further research, which is necessary for a more comprehensive understanding of new organizational designs. Many key aspects of these are still little understood.

Some necessary conditions for the existence of these new organizational designs enjoy agreement in the literature, such as smaller size (Carter, 1984; Galbraith & Lawler, 1993) and more dynamic organizational environments (Burns & Stalker, 1961; Powell, 1990); others are controversial, such as the role of strategy (see Baker, 1992; Mintzberg & McHugh, 1985). Although normative prescriptions regarding the IT importance loom large (e.g., Benjamin & Scott Morton, 1992; Fulk & DeSanctis, 1995; Handy, 1989; Morgan, 1993; Rockart & Short, 1991), they remain speculative without provision of appropriate evidence. The literature is inconclusive as to whether IT is a fundamental pillar, a buttress, a lintel, or even just a design ornament resulting from organizational mimicry. Confusion, furthermore, lurks in the light of common experience that non-bureaucratic organizations can exist with a small use of IT (e.g., small start-ups), while bureaucracies can have abundant IT (e.g., military organizations). The lack of holistic studies, furthermore, prevents one from differentiating at a large picture-level between the new designs and their antipode—the bureaucracy, which also can be based on information, knowledge, and IT. Differentiation between the information aspects of the two antithetical organizational designs has, therefore, to be called into question. This is even more urgent provided that the bureaucracy is far from being at the fringe of history, nor is it stripped from accolades (e.g., Ashkenas, Urlich, Jick, & Kerr, 1995; Jaques, 1989).

Given our meager knowledge of the information aspects of the new organizational designs, the present study intends to contribute to both the normative and empirical domain of these. In realizing this goal, the study (1) proposes a model of a non-bureaucratic organizational design, which is based on relevant literature and will be called the non-traditional organization; and (2) provides a preliminary empirical test of this model.

## 2. Previous Research

### 2.1. Information Technology

For the purposes of the present study, IT is conceptualized as electronic machines, devices, and their applications that have both computing and communication capabilities. The concept has support in relevant literature. For example, sociologist of organizations, Child (1987), defines IT as technologies and applications which combine the data-processing and storage powers of computers with the distance-transmission capabilities of telecommunications. Similarly, management researcher Huber (1990) defines "advanced IT" as devices (a) that transmit, manipulate, analyze, or

exploit information, and (b) in which a digital computer processes information integral to the user's communication or decision task. This technology is also called "communication" or "telecommunications technology," "information and communication technology," "information/communication technology," etc. Instances of IT investigated in the present study are E-mail, electronic bulletin boards, electronic file transfer, shared electronic databases, the facsimile (fax), voice mail, and the telephone. The last three belong to important communication technologies which exhibit smaller but increasing computing capabilities (the core of voice mail systems is a switching computer, the fax can be mounted on a computer, while the telephone is being used for applications in computer-telephone integration).

The communication capabilities of IT (sending, receiving, forwarding, narrow- and broadcasting, etc.) make it possible for organization members to exchange work information (opinions, questions, advice), design the flow of work (e.g., routing and management of electronic documents), and to socialize (e.g., through chat-sort of electronic bulletin boards). The computing capabilities of IT (storing, retrieving, processing, etc.) provide the tools for transforming work material (e.g., processing data for analytical purposes). The blend of these two capabilities has a history with various organizational implications. For example, E-mail was initially an interpersonal communication medium for overcoming the telephone tag problem, with computing capabilities limited to processing and storing of electronic text (cf. Rice, 1984). Later on, other capabilities were added, such as forms support, filtering, and retrieval that exist, for example, in Winograd's (1988) Coordinator. These transformed E-mail into technology for supporting work groups (groupware). Groupware could then be used for organizing work in team fashion, with all the organizational implications that follow. But this is true only in principle.

The ontology of the IT role in the organizations is not simple. IT can generate opportunities for new methods of organizing people, work, material resources, and time. But, whether these opportunities will be perceived and used depends on a particular organizational context (see Ciborra, 1996). The organizational context can make use of technological capabilities, or restrict their use, or even invent new uses reaching beyond these capabilities. The multi- and mutually causal relationships between the organizational and the technical refer to an epistemology that can be identified in the socio-technical and the emergent process (Markus, 1994) approaches, as well as in the social informatics field of study (see Kling, 1997).

The relationships between IT and specific organizational aspects have been investigated empirically. One such aspect is organizational structure (patterns of stable social relationships), which can be sufficiently explained by the dimensions of hierarchy, centralization, and formalization (Fry, 1982). Hierarchy was investigated by Whisler (1970), who found that a reduction of hierarchy layers was associated with the use of computers in the insurance industry. Similarly, Pool (1983) argued that the telephone led to aberrations in the hierarchical patterns of communication in the old steel industry because it made it possible for workmen to directly access executives. In the case of formalization, the literature is inconclusive. While Pfeffer and Leblebici (1977) found, contrary to the their expectations, that the utilization of IT in small factories was not associated with the use of more formal written reviews or statistics on departmental performance, Wijnhoven and Wassenaar (1990) found in their case studies that communication applications at a high-tech factory and at a computer service firm led to more formalized communications. Normative theory does not ease the dilemma. For example, McKenney's 1985 study (as cited in Wijnhoven & Wassenaar, 1990) assumes that IT-supported communication seems to facilitate more organic organization, by allowing easy and timely linking to a broad set of individuals with common referent bases (see Wijnhoven & Wassenaar, 1990, p. 51). On the contrary, Huber (1984) posits that IT can trigger requirements for formalizing the handling of non-routine information, scope and amount of environmental information, and decision making; in his later theorizing, however, Huber (1990) hypothesizes that long-term effects of IT on formalization associated to it could be nil: When IT is viewed from the perspective of a life cycle, the amount of rules regulating the use of a system can be high in early stages and can decrease as the system becomes familiar, approaching the degree of formalization associated with the old, replaced system. This neutral role of IT is in agreement with the finding of Pfeffer and Leblebici (1977).

Centralization, another crucial structural dimension, was also found to be associated with IT, albeit in a diversified fashion, depending on the level of management and organizational size. Specifically, at the operational level, a link between IT and a decrease in centralization was discovered in railroad management (Dawson & McLaughlin, 1986) and in city departments of human resources (Keon, Vazzana, & Slocombe, 1992). At the executive level, however, IT was found to be related to increased centralization in the insurance industry (Baker, 1992), large newspaper organizations (Carter, 1984), and railroad management (Dawson & McLaughlin, 1986). Smaller organizations appear to provide the context in which IT is associated with a decrease in overall centralization, specifically in factories (Pfeffer & Leblebici, 1977), newspaper organizations (Carter, 1984), and in units of a hospital (Barley, 1990). This evidence suggests that the relationship between a decrease in centralization and IT use is likely to exist at the operational level of an enterprise and/or in smaller organizations. Finally, the relationship between IT use and the spatial dispersion of operations and organization members was discovered in organizations of scientists (Hiltz, 1984; Kerr & Hiltz, 1982), a software company (Olson & Bly, 1991), and other organizations (Sproul & Kiesler, 1991).

Organizational culture (values and patterns of behavior shared among organization members) was also investigated in relation with IT. Several studies in laboratory and orga-

| Structural effects | Study |
|---|---|
| Decrease in hierarchy | Pool (1983); Whisler (1970) |
| Centralization strategic | Pool (1983); Whisler (1970); Carter (1984; large firms), Dawson & McLaughlin (1986); Wijnhoven & Wassenaar (1990) |
| Decentralization overall | Pfeffer & Leblebici (1977); Carter (1984); Barley (1990); Wijnhoven & Wassenaar (1990) |
| Decentralization operational | Dawson & McLaughlin (1986); Keon et al. (1992) |
| Decrease in formalization | McKenney (1985) (as cited in Wijnhoven & Wassenaar, 1990); Pfeffer & Leblebici (1977) |
| Increase in formalization | Huber (1984); Wijnhoven & Wassenaar (1990) |
| Facilitating spatial dispersion | Olson & Bly (1991); Sproul & Kiesler (1991) |
| **Cultural effects** | **Study** |
| Uninhibited behavior, social equalization, decision shifts, new ideas creation | Kiesler et al. (1984); Sproul & Kiesler (1991) |
| Conformity avoidance | Smilowitz et al. (1988) |
| Expression of feelings | Rice & Love (1987); Wigan (1991) |
| Community creation | Foulger (1990); Kerr & Hiltz (1982) |
| Knowledge sharing | Dhar & Olson (1989); Olson & Bly (1991) |
| Accountability | Guttiker & Gutek (1988) |
| Cross-departmental communication | Kerr & Hiltz (1982); Hiltz (1984); Shapiro & Anderson (1985); Olson '& Bly (1991); Sproul & Kiesler (1991) |
| **Political effects** | **Study** |
| Salience of expert power | Burkhardt & Brass (1990); Barley (1990) |

FIG. 1. Evidence of the role of information technology in organizations.

nizational contexts found that E-mail and other IT, due to a lack of the social cues they exhibit, could lead to relatively uninhibited behavior, social equalization, decision shifts, and the creation of new ideas (Siegel, Dubrovsky, Kiesler, & McGuire, 1986; Sproul & Kiesler, 1991). This somewhat uninhibited behavior was sometimes interpreted as an undesirable aberration from social norms; however, there was no agreement as to whether this was attributable to IT (Kiesler, Siegel, & McGuire, 1984) or to its users (Smolensky, Carmody, & Halcomb, 1990). This social aberration was sometimes deemed stimulating for organizational innovation (Siegel et al., 1986; Sproul & Kiesler, 1991). Innovation occurrences, in turn, could be brought into relation with studies of cooperation and knowledge sharing in IT-rich organizations (Dhar & Olson, 1989; Olson & Bly, 1991), conformity avoidance (Smilowitz, Compton, & Flint, 1988), expression of feelings (Rice & Love, 1987; also see Wigan, 1991), and the creation of an electronic community within professional organizations (Foulger, 1990; Kerr & Hiltz, 1982).

Lastly, IT has been studied from the perspective of expert power—influencing others upon possession of special knowledge, a capability to deal with uncertainty, and an image of being irreplaceable (cf. Mintzberg, 1983; Pfeffer, 1981; Tushman & Anderson, 1986). It was discovered that the knowledge of using IT generates power over those with no such knowledge (Barley, 1990; Burkhardt & Brass, 1990). Figure 1 summarizes the empirical research on IT and the organizational dimensions discussed above.

### 2.2. Non-Bureaucratic Organizational Designs

Organic organization, adhocracy, and the network organization are three designs that played a pivotal role in conceptualizing the model of the non-traditional organization in the present study. The attention these designs have received, as well as their level of elaboration, led to selecting them over other designs. Each is discussed below.

Burns and Stalker's (1961) *organic organization* is a historical blueprint for new organizational designs. This label is supposed to signify the spontaneous, "natural" genesis of an organization, one that resembles Durkheimian juxtaposing of the "organic" to the "mechanistic" principle. The authors derived this design from their empirical studies of successful electronics firms. The organic organization is characterized importantly by work organization resting on flexible tasks; control and authority that are distributed in a network fashion; and a culture recognizable after a "technological ethos" rather than loyalty and obedience. The model does not account for IT, but does include some other information dimensions—communication content, specifically, which should be information and advice rather than instructions and decisions.

Mintzberg's (1979) *adhocracy* has been another influential design, which can be traced back to the work of Toffler (1980) and Bennis and Slater (1968). The term comes from Latin roots which convey that immediate purposes shape an organization. Mintzberg's adhocracy exhibits low formalization, a flatter hierarchy, vague roles, the reliance on

specialized teams that become main power holders, employee-cooperation-intensive communication, and democratic values. In his power studies, Mintzberg (1983) posits that power accumulates in the hands of experts who carry out complex, specialized work, and possess knowledge acquired in professional societies and universities rather than firms; this power is not rooted in formal authority, but in self-coordination, teamwork, and mutual adjustment in adhocratic organizations. Moreover, Olson and Bly (1991) provide evidence of a *distributed organization* experiment which took place at Xerox. The authors discovered that organization that developed at two sites, which were connected via a multimedia system, was closest to Mintzberg's adhocracy (flat, peer-based control, ambiguous roles) and devised several interesting information dimensions. Specifically, information was the primary organizational resource; the information system linking and supporting the two work sites was necessary for this organization to function; informal communication was essential to information sharing, coordination, and socialization; communication was intensive, extended to the social domain, and made the borders between private and public turfs fuzzy; a positive relationship developed between the extent of information sharing and the degree of reciprocal interdependence (each party possesses knowledge/information that others need and needs knowledge/information in possession of others). Another interesting study was conducted by Mintzberg and McHugh (1985). They found that the National Film Board of Canada metamorphosed from a bureaucracy into an adhocracy, owing to successive attempts at redistributing power to professionals. This long process resembled "the growth of garden weeds," emerging through complicated and hardly traceable organizational processes. Bailey and Neilsen (1992), however, discovered that adhocracy could exist in parallel with bureaucracy in an instructional unit within a large corporation they studied. The adhocracy was exemplified in undifferentiated leadership, and the ability to deliver innovative services while working under an ambivalent mission statement; the bureaucracy, on the other hand, was perpetuated though the steady delivery of standardized services.

The third interesting design is the *network* or *networked organization.* The term is used both with reference to inter-organizational designs (Aldrich, 1981; Myles & Snow, 1986) and intra-organizational ones, which are of interest here. Rockart and Short (1991) have proposed a design of the networked organization, at the nexus of which are informal human networks enabled by IT networks. These socio-technical networks can foster the redesigning of the division of labor, and the development of a new culture characterized by the sharing of goals, expertise, decision making, recognition and reward, responsibility, accountability and trust. Computer-supported social networks are also important in Morgan's (1993) loosely-coupled organic network, which rests on a founding/driving team and subcontracting, and resembles a spider plant—umbilical cords are embedded in information systems, accountability, obligations, shared resources, visions and values. In this design, IT enables organization members to be "completely connected while remaining far apart." In empirical investigation, Eccles and Crane (1987) have found that some American and foreign commercial banks resemble very flexible networks (e.g., dual reporting, vague roles, constant flux); one characteristic of these networks is the "substantial potential for conflict." Similarly, Quinn (1992) portrays the crucial role of IT in Arthur Andersen and Company, one of the "Big Six" public accounting firms which the author considers an instance of the network organization. An intensive sharing of information and knowledge among local offices characterizes the firm, which is enhanced by incentive systems and organizational culture. Baker (1992) defined the networked organization as one which is highly integrated vertically (hierarchical levels), horizontally (functional division of labor), and geographically (spatial spread of organizational units). He went on to test this definition on a small real estate firm, which, in fact, was driven by a network organizational strategy, and found that the firm matched his definition only moderately well.

Elements of broader theorizing about new organizational designs can also be found in the literature. Heydebrand (1989) portrays the new designs as *smaller size* organizations that are in either the information or service business, have computerized technology, and employ informal and flexible division of labor. The management in these organizations is functionally decentralized, eclectic and participative, while the control structure is intermingled with pre-bureaucratic elements, such as clan-like personalism and informalism. He hypothesizes that computer technology plays an essential role in the service sector, because it makes it possible for internalizing of formal rationality, calculability, and procedures into software and systems on which production is organized. A consequence of this internalization is the development of informal social and work relationships, including a minimal division of labor, de-hierarchization, and only general rules to permit flexible social control and self-regulation. The technocratic ratio displaces the bureaucratic, concludes Hedeybrand. IT and information-related expertise are also key factors in Galbraith and Lawler's (1993) conceptualization of new designs. Indeed, these designs are defined by a capability of managing and taking advantage of the "new world of information complexity and richness": The challenge is in creating technology and the individuals who understand the importance of information and know how to access and use it. These designs, furthermore, tend to be *decentralized,* use an internal network of divisions and units linked through electronic forms of communication (he terms this "distributed organization"), diminish the importance of hierarchy so that it becomes just one among many of coordination and control vehicles, and possess adaptation capabilities which result in heterogeneous organizational methods within the same firm ("fluid and transitory design").

Zuboff (1988) reminds us that IT can be used for either petrifying or undermining hierarchical authority, a hallmark

of bureaucracy. If IT is used just to automate the existing division of labor, power structure will remain intact and no change will occur in the distribution of learning or access to information. IT can, however, be used to change the division of labor, which could result in power redistribution down the ladder, and different patterns of information access and learning—a new organizational design that Zuboff terms *informated organization.* This organization relies on the human capacity for teaching and learning, on criticism and insight, and on deploying both human and technology capabilities to their full potential. Moreover, Clegg (1990, 1996) places new organizational designs into the realm of postmodern. Drawing on European contemporary tradition, Clegg (1990) suggests that the phenomenon of "de-differentiation," which was suggested by Lash (1988) as a criterion of separating postmodern culture from the modern one, applies to postmodern organizations as well. De-differentiation characterizes a multitude of organizational aspects, such as a broadening of work roles, turning accountability relationships from individual ones into team-based ones, expanding narrow expertise into flexible specialization, and replacing employment relations by complex relational forms of networking and subcontracting. De-differentiation comes from the demands for product differentiation characterizing today's marketplace.

Taylor and Van Every (1993) propose that the relationship between IT and organizations could be approached from the perspective of juxtaposing metaphors of *text and dialogue.* These can help understand successes or failures in the process of designing both computer systems and the organizational context. Whereas text induces rather inflexible, rules-driven system development and organizational contexts, dialogue resembles ongoing, live, action-rich development processes and organizational contexts. Where text rules over dialogue, office automation projects are likely to fail. In a similar vein, Heckscher (1994) elevates dialogue to a cornerstone of his framework of *interactive organization.* Dialogue is the vehicle for developing principles of action that replaces traditional rules and regulations, facilitates the creation and maintenance of shared organizational vision and strategies, and underpins the sharing of trust; trust, in turn, is the basis of persuasive ability that substitutes for power relationships, and it helps legitimize the necessary hierarchy, concludes Heckscher. Because dialogue and text carry connotations of information artifacts, they have an intrinsic appeal in understanding new designs from the information perspective.

### *2.3. Public Accounting Industry*

The segment of the public accounting industry (accounting industry) investigated in the present study, the so-called Big Six, is engaged in three businesses—account auditing, tax preparation and planning, and consulting. In Stevens's (1981, 1985) discussion on history of this industry, account auditing features as the oldest business. It was responsible for starting the modern accounting industry in the U.S. in the late 19th century—an import from Scotland intended to strengthen financial control of British investors over their increasing stakes in the U.S. economy. The tax business was added in the early 20th century in conjunction with modern tax laws. The industry started providing consulting services in the second half of this century, first as "advisory services" for financial management, then on a broader scale to include design and management of information systems. Within the tax business, the accounting industry faces competition from legal firms, while software/hardware vendors are the competitors within the consulting business.

Accounting firms represent an interesting organizational context. Their legal basis is partnership, which is capable of creating variation within certain contextual constraints. This is particularly true of the local office, which is the basic operating unit organized on territorial, customer, and functional principles. For example, management incentives for using IT can be unique to a local office, while IT strategy applies across local offices. An implication is that the local office represents the scene of interesting interaction between structure and action. Moreover, production in the local office is based on group work. This work is thus carried out by the "engagement team" (team) which directly serves a particular client. Teams report to a particular manager or partner (the partner is both the partial owner and the executive), while maintaining a certain discretion in organizing their work, managing the budget, and so on. Teamwork, adopted in business organizations as late as the 1980s, has, in fact, been a hallmark of the modern accounting industry since its inception. Another interesting characteristic of the local office is typically a small size (up to 500 employees), which resolves the well-known problem of modifying effects of a larger organizational size.

The local accounting office is an information business par excellence. Its production is comprised of data collection and processing, information creation, and knowledge development. All this demands extensive capabilities in information management (data collection, storing, retrieving, checking, comparing, estimating, transforming) as well as copious knowledge (as an executive put it, "a consultant has to understand organization, planning, performance aspects, measurement, accounting, reporting issues, and the development and analysis of information for management use" (Stevens, 1981, p. 119). Informal communication facilitates sharing of information and knowledge in audit tasks (Dirsmith & Covaleski, 1985). Moreover, the Big Six segment firms were early adopters of IT and remain heavy users (Sutton, 1992), which also suits the research problem of the present study.

The tension between the organic and the bureaucratic has been described in the literature of organization of the accounting industry. Sorensen (1967) and Sorensen and Sorensen (1974) found that senior organization members (partner, manager, supervisor, and senior accountant) tended to enforce more bureaucratic values and behavior, while the junior organization members (junior or staff accountant) exhibited more "professional" values and behavior. Pratt

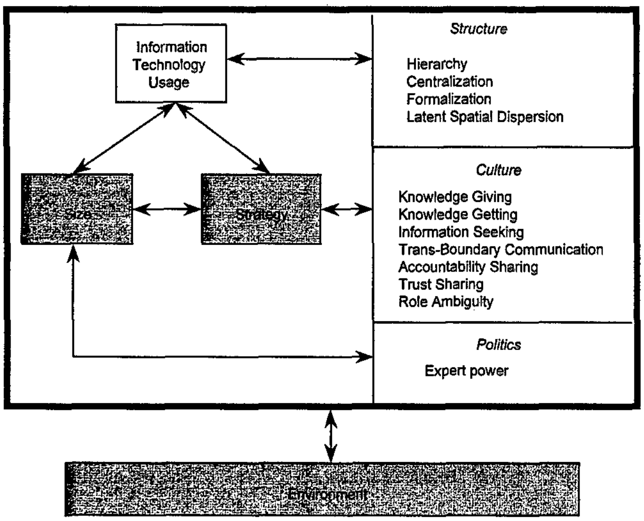


FIG. 2. Research model. Note: The large rectangle represents organization. Variables in shaded boxes are controlled. Order of variables is arbitrary.

and Jiambalvo (1981) studied performance of audit teams, and found that both the organic and bureaucratic modes of organizing could emerge as a consequence of the team leader's behavior. Watson (1975) found that the alternatives in organizing depend on different levels of uncertainty in auditing versus consulting tasks, and that the lower uncertainty in auditing renders it more bureaucratic, while the opposite is true of consulting. However, Ballew's (1982) study established no support of these findings—significant differences existed neither in the organization of different practice teams (vertical distribution of power, centralization, and formalization) nor in two of the three dimensions of the technological routines he investigated. Approaching design of accounting organizations from a somewhat different perspective, Stevens (1985) and Weinstein (1987) provided visions of the industry's leaders concerning the evolution of accounting firms from pyramidal structures into flatter, market-responsive organizations. Similarly, Quinn (1992) described the network characteristics of a Big Six firm that surfaced in the management of information and knowledge, deployment of IT, daily operations, and organizational structure. Evidence of the characteristics of new organizational designs in the accounting industry is, therefore, inconclusive and incomplete, but certainly conducive to further research.

## 3. Methodology

This section discusses the research model, research questions, measurement, pilot investigation, units of data analysis and data collection, and data collection and data analysis methods.

The present study was guided by the research model depicted in Figure 2. The research model contains:

- Structural dimensions of formalization, centralization, hierarchy, and latent spatial dispersion;
- information dimensions of IT usage, knowledge getting, and knowledge giving (these last two can also be called cultural variables);
- cultural variables of accountability sharing, trust sharing, and role ambiguity that can additionally explain team-based organization;
- demographic variables of age, years with company, and years in field;
- an expert power dimension; and
- control variables of size, strategy, and environment.

The model juxtaposes IT usage with other variables in order to respond to the first research question (see below). Symmetric relationships among variables are presumed, which is in accord both with the state of theorizing on new designs and the cross-sectional character of the present study. It cannot be overstated that the selection of variables is just one of many possibilities, which builds on emphases in the relevant contemporary literature and on parsimony concerns. Definitions of the variables, along with the corresponding measures, validity and reliability estimates, appear in Table 1. Scales for measuring variables (four- or three-item composites) are available from the author upon request.

These research questions guide the present study:

1) What role does IT play in the non-traditional organization.
2) What are the other important characteristics of the non-traditional organization.

In order to obtain quantitative information for answering the first question, two index measures were designed—one representing the measures of IT used, and the other representing a compound of other variables studied (with the exception of control variables). In effect, each of the 12 local offices studied was represented with one IT index and one NT index (see Tables 1 and 3). The IT investigated was operationalized as the amount of usage of electronic mail, electronic bulletin boards, electronic file transfer, shared electronic databases, the facsimile, voice mail, and the telephone.

The second research question reflects the purpose of the present study to shed light on the relationship between information dimensions and other relevant organizational dimensions in new organizational designs. Part of this understanding is certainly the answer to the first question; another part is the knowledge that can accrue from looking more closely at the variables comprising the model of the non-traditional organization the study has introduced and tested.

A pilot investigation set out to define the study population. The literature pointed to the accounting industry as a candidate (see the discussion above), and this led to a pilot investigation of small accounting, insurance, engineering, data services, and manufacturing firms in the residential area of a typical mid-size city on the East Coast. Investigated were the dimensions of hierarchy, departmention,

Table 1. Variables and measures.†

| Construct | Measure | Validity | Reliability |
|---|---|---|---|
| 1. **IT usage**—The extent to which IT is used (Hiltz, 1984) | 1.1. The frequency of usage within time periods (Davis, 1989) | $r = 0.63$ | Cronbach $\alpha = 0.81$ |
| | 1.2. The number of transactions via IT in a typical day (Adams, Nelson, & Todd, 1992) | $r = 0.92$ | Cronbach $\alpha = 0.81$ |
| 2. **Formalization**—The extent to which rules, procedures, instructions, and communications are written (Oldham & Hackman, 1981) | The amount of written rules and procedures (Oldham & Hackman, 1981) | Discriminant, among 4 structural variables $r_s = 0.11$–$0.25$ | Cronbach $\alpha = 0.52$ |
| 3. **Centralization**—The extent to which control is concentrated (Tanenbaum, 1961) | The amount of control held by managers (Markham, Bonjean, & Corder, 1984) | N/A | Split-half coefficient 0.39 |
| 4. **Hierarchy**—The extent to which levels of graded authority exist (Weber, 1946) | The number of management levels; new measure | N/A | N/A |
| 5. **Latent spatial dispersion**—The extent to which is possible to perform joint work at a distance (Hiltz, 1984) | The estimate of the importance of physical proximity for getting the job done; new measure | N/A | Cronbach $\alpha = 0.73$ (N = 29)* |
| 6. **Information gathering collaboration**—The extent to which a person relies on others in gathering work-related information (Olson & Bly, 1991) | The frequency of asking co-workers to provide information needed; new measure | N/A | Cronbach $\alpha = 0.53$ (N = 29) |
| 7. **Knowledge getting**—The extent to which a person gets knowledge from colleagues (Rockart & Short, 1991) | The frequency of one's getting knowledge from his co-workers and other sources; new measure | N/A | Cronbach $\alpha = 0.72$ (N = 29) |
| 8. **Knowledge giving**—The extent to which a person makes his knowledge available to co-workers (Rockart & Short, 1991) | The intensity of one's attitudes toward giving knowledge to his co-workers; new measure | N/A | Cronbach $\alpha = 0.53$ (N = 29) |
| 9. **Trust sharing**—The extent of a team member's expectancy that statements of his co-workers can be relied upon (MacDonald, Kessel, & Fuller, 1972; Rotter, 1967) | The intensity of one's attitudes toward reliance and faith in his colleagues and people in general (MacDonald et al., 1972) | $r = 0.56$ | Cronbach $\alpha = 0.84$ |
| 10. **Accountability sharing**—The extent of a team member's condition of being ready to justify his team's decisions to superiors (Tetlock, 1983) | The intensity of one's readiness to share the burden of justification with his team members; new measure | N/A | Cronbach $\alpha = 0.77$ (N = 29) |
| 11. **Trans-boundary communication**—The extent to which a person communicates across his team's boundaries (Hiltz, 1984) | The frequency of communications a team member makes with people out of the team; new measure | N/A | Cronbach $\alpha = 0.70$ (N = 29) |
| 12. **Expert power**—The extent to which a person is influenced by knowledge holders, so that he does things he would not have done otherwise (Dahl, 1957; French & Raven, 1959) | The estimate of the expertise and other sources of influence impinging on a person (Bachman, Bowers, & Marcus, 1968); modified scale | N/A | N/A |
| 13. **Role ambiguity**—The extent to which necessary information is lacking to an organizational position (Kahn, Wolfe, Quinn, Snoek, & Rosenthal, 1964) | The intensity of one's perception that necessary information is available (Rizzo, House, & Lirtzman, 1970) | Concordance of statistics across samples = 0.84 | Convergent reliability 0.88 |
| 14. **Environment**—The extent to which forces outside of organizational boundaries affect an organization (Lawrence & Lorsch, 1968) | The assessment of clients, and competitors a local office deals with; new measure | N/A | N/A |
| 15. **Strategy**—The extent to which an organization's plans of long-term goals affect the organization (Chandler, 1962) | The description of a local office's business and IT strategies; new measure | N/A | N/A |
| 16. **Size**—The scale of operations in an organization (Price & Mueller, 1986) | The number of full-time employees; new question | N/A | N/A |

Table 1. (continued)

| Construct | Measure | Validity | Reliability |
|---|---|---|---|
| **17. IT index**—An office's average use of IT | Compound measure building on measures 1.1 and 1.2 $$\text{IT Index} = \frac{\sum_{j=1}^{n_{ij}}(\text{ITFRE}+\text{ITTRA})}{n_i}$$ j = respondent, i = local office, ni = the number of respondents in a local office; new measure. See computation note | Dependent on the validity of belonging measures | Cronbach $\alpha = 0.64$ (N = 131) |
| **18. NTO index**—An office's average of 12 dimensions of structure, culture, and politics | Compound measure building on measures 2 through 13 $$\text{NTO Index} = \frac{\sum_{j=1}^{N_{ij}} Yk}{n_i}$$ $j$ = respondent, $i$ = local office, $n_i$ = the number of respondents in $i$-th local office, $Yk$ = $k$-th dimension belonging to the index; new measure. See computation note | Dependent on the validity of belonging measures | Cronbach $\alpha = 0.36$ (N = 116) |
| **19. Age** | One of six categories corresponding to the respondent's age (under 20, 20–29, 30–39, 40–49, 50–59, 60 and over) | N/A | N/A |
| **20. Years with company** | The number of years spent with the current company | N/A | N/A |
| **21. Years in field** | The number of years spent in tax, auditing, or consulting practice | N/A | N/A |

[†] Computation note: The IT index: (1) Standardize raw scores of both IT usage measures; (2) sum the standardized scores for each respondent; (3) sum the result from step 2 across individuals; and (4) divide the result from step 3 by the number of respondents in a given local office. The NT index: (1) Code inversely original scores of structural variables reflecting the traditional organizational structure (hierarchy, formalization, and centralization); (2) standardize raw scores of all dependent measures; (3) sum the standardized scores of all dependent measures for each respondent; (4) sum results from step 3 across respondents; and (5) divide the result from step 4 by the number of respondents in a given local office.

* Estimated on an initial sample of 29 respondents from three local offices (later included in the study sample). Scale calibration combined analyses of covariance-based Cronbach alphas, Pearson's correlation coefficients, and factor analysis—orthogonal rotation by the varimax method.

span of control, formalization, team work, management promotion policies, pace of organizational change, and the IT profile (kinds used, the extent of use, perceived effects of the use, and perceived importance). The main finding was that accounting firms (local offices) ranked first on the dimensions of team work and of the IT profile; in terms of structural and management properties, these offices also appeared to represent desirable characteristics. This effectively confirmed the literature-based expectations, and promoted the accounting industry into the study population.

The present study defined "organization" as the local accounting office, because of its unique, suitable characteristics discussed in the previous section. The sample studied consisted of 12 such organizations located on the East Coast and in the Midwest from five of the Big Six accounting firms. The original intention of drawing a random sample of the offices was only partially realized due to the survey unavailability of some offices. In each office, however, a random stratified sample of the professional staff was taken and surveyed extensively on all but the strategy and environment dimensions, which were investigated by interviewing an executive (managing partner, partner, or manager) in each office. Characteristics of samples appear in Table 2.

The data collected were analyzed by correlation analysis (bi- and multivariate), factor analysis (principal component—varimax rotation), multiple regression analysis, and content analysis. The quantitative data were first explored in terms of distribution of the variables; this analysis indicated no properties that would require transformation of data.

In order to answer the first research question about the relationship between IT and the non-traditional organizational dimensions, the IT indexes and NT indexes for the 12 organizations studied were correlated, while controlling for organizational size and environment (strategy data, which were qualitative, were content analyzed, and no significant differences among the offices were found). Then, correlation analysis of zero-order was conducted as the start point in exploring relationships between all the variables in the model (the second research question). Factor analysis was conducted in turn, in order to determine sets of variables within the relatively complex model of the non-traditional organization. The next step was an intricate partial correlation analysis of the variables within the sets and at their borders. The goal of this analysis was to determine moderating variables. In order to see, for example, if IT relates with variables that appear entirely detached in zero-order

Table 2. Characteristics of samples.†

| Local office | Ns | Nt | Ps | N | A | T | C | Pr |
|---|---|---|---|---|---|---|---|---|
| A | 8 | 84 | 10 | 170 | 7 | 1 | 0 | 87/13 |
| B | 15 | 200 | 8 | 410 | 9 | 4 | 2 | 60/40 |
| C | 19 | 125 | 15 | 238 | 16 | 1 | 2 | 84/16 |
| D | 8 | 55 | 15 | 100 | 4 | 3 | 1 | 50/50 |
| E | 7 | 50 | 15 | 97 | 5 | 2 | 0 | 71/29 |
| F | 16 | 116 | 14 | 160 | 1 | 5 | 10 | 6/94 |
| G | 8 | 50 | 17 | 65 | 5 | 3 | 0 | 63/37 |
| H | 13 | 170 | 8 | 300 | 9 | 2 | 2 | 69/31 |
| I | 8 | 58 | 16 | 63 | 8 | 0 | 0 | 100/0 |
| J | 10 | 120 | 8 | 120 | 8 | 2 | 0 | 80/20 |
| K | 4* | 13* | 30 | 130 | 0 | 0 | 4 | 0/100 |
| L | 15 | 85 | 18 | 140 | 12 | 3 | 0 | 80/20 |
| Total | 131 | — | — | — | 84 | 26 | 21 | 65/35 |

† *Ns* = number of respondents in an office; *Nt* = number of professional staff in an office; *Ps* = percentage of professional staff surveyed in an office; *N* = total number of employees in an office; * = consultants only; *A* = auditors; *T* = tax experts; *C* = consultants; *Pr* = percentage ratio of auditors vs. tax experts and consultants combined.

correlations, complete paths including these variables were tested by means of partial correlations. Furthermore, multiple linear regression analysis (the stepwise solution) on all variables that appeared important, either on the basis of theory or the previous analysis, was then conducted. Regression analysis was also used for testing the data for collinearity—tests of tolerance, variance inflation factor, condition index, and residual plots. No collinearity was identified. The quantitative methods employed were deemed appropriate given the cross-sectional character of the data and the size of samples.

## 4. Findings

The partial Pearson's correlation coefficient measuring the relationship between the IT index and the NT index is 0.68 (see Table 3). This could be considered a strong association, indicating that a higher usage of IT and non-traditional dimensions coincide in the same context. Further analysis of data on the entire pool of respondents across all offices ($N = 131$) reveals that IT usage correlates negatively with formalization and centralization, and positively with trans-boundary communication, trust sharing, and demographic variables. Figures 3 and 4 depicts these and other relationships which represent the state of the non-traditional organization that has been found to exist in the sample studied. These figures use the variables from the research model of the non-traditional organization (Fig. 2) and establish their relationships on the basis of correlation and factor analysis.

Analysis of all the seven relationships in which IT usage participates has revealed that the relationship between IT usage and formalization is robust, while all others are moderated. This specifically applies to relationships within the set consisted of IT usage, trans-boundary communication, centralization, and trust sharing. This set rests on an underlying dimension, which will be referred to as IT-communication factor (see Table 4). Partial correlation analysis indicates that IT relationships within this set are moderated by centralization and trans-boundary communication (see Table 5). (For convenience, a moderating effect is considered one that alters the zero-order correlation by at least 5% of the variance explained, or an effect that changes the level of significance regardless of meeting the 5% criterion, e.g., when the $p$ would take on worse than the 0.5 value.) On the other hand, IT usage amplifies the strength of the relationship between trust sharing and trans-boundary communication.

Another piece of evidence of the role of IT in the non-traditional organization concerns its predictive power. Multiple linear regression analysis has revealed that IT usage is the best predictor of decentralization and trans-boundary communication, and the second best predictor of trust sharing and formalization (Table 6). This makes IT usage the variable with the greatest predictive power among all the variables investigated.

Part of the explanation of the role of IT in the non-traditional organization refers to qualitative valuations of office technology. The respondents were asked to name the technology which they could not imagine their work without, whichever this happens to be—a pen, a manual, computer program, anything. Answers suggest that the computer is that technology, while the telephone and voice mail come in a close second (see Table 7). The importance of the computer hinges on its qualities of being the basic work tool (e.g., spreadsheets) and a main telecommunication channel (e.g., E-mail and the facsimile). The most interesting comments associated the computer with freedom from the traditional constraints of place, structure, and workflow, that is, with its potentials related to supporting new organizational designs:

> With applications on my PC system and access to a phone jack, I can set up my office anywhere.
>
> I use my laptop several times each day for document production, storage, and document transmission. I can virtually set up my office anywhere with it.

Table 3. Statistics of the NT index and the IT index ($N = 12$).*

| | *N* | Mean | *SD* | Max | Min | Range |
|---|---|---|---|---|---|---|
| NT index | 12 | 58.84 | 0.62 | 60.24 | 57.95 | 2.29 |
| IT index | 12 | 60.02 | 0.91 | 61.49 | 58.54 | 2.95 |
| Correlation between IT index and NT index | | | | | | |
| Partial $r$ (size and environment removed) | | | | | | $0.68, p < 0.03$ |
| Spearman rho | | | | | | $0.65, p < 0.02$ |

* The indexes are standardized scores with 60 added to each to make them more readable. All values rounded to the second decimal point. *N* = sample size; *SD* = standard deviation; Max = maximum score; Min = minimum score.

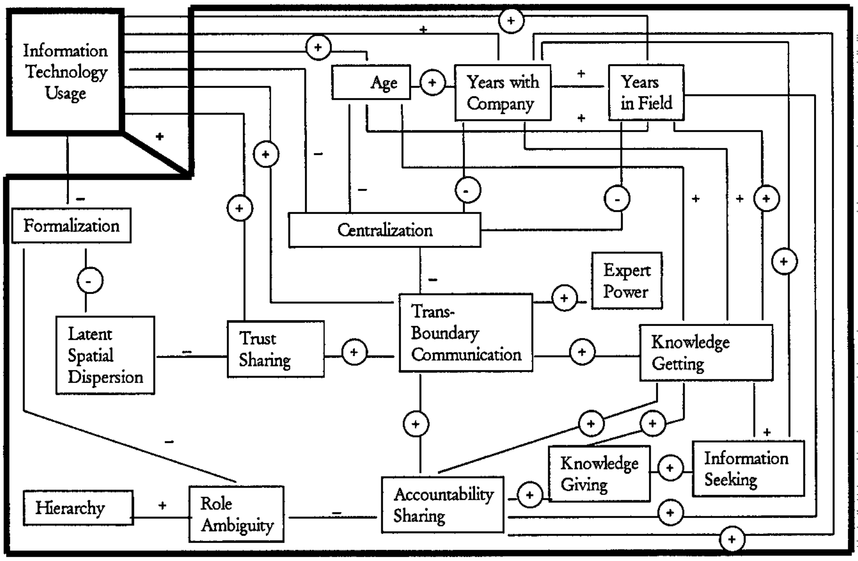


FIG. 3. Findings on the non-traditional organization. Note: Thin lines represent bi-directional relationships based on significant partial correlations, here designated by their direction. A direction sign appearing within a small circle stands for a relationship that depends on other variables (the Pearson $r$ either loses or acquires significance with partialling out relevant variables). IT usage is connected with the rest of the organizational context by a solid line in order to represent the relationship between the IT index and the NT index.

In answering the second research question, which addresses other important characteristics of the non-traditional organization, several findings are relevant. First, two variables from the IT-communication set deserve attention—trans-boundary communication and trust sharing. Trans-boundary communication (interactions beyond team boundaries) appears to be a significant linker in organizational design: It is related to technology (IT usage), structure (centralization; a negative relationship), politics (expert power), other information aspects (knowledge getting), work organization and management (accountability sharing), and a psycho-social domain of organizational culture (trust sharing). A detailed partial correlation analysis suggests, however, that trans-boundary communication does not moderate relationships between these variables. Trust sharing, on the other hand, is one of the much-debated issues in the literature of new organizational designs. Its importance in the organizations studied may be indicated by the fact that it also participates in another variable set called Role Ambiguity (Table 4). Looking from the perspective of new organizational designs, the discovered direct relationship between trust sharing and IT usage is a desired characteristic. However, this characteristic is somewhat offset by the finding that trust sharing is best predicted from inversely related latent spatial dispersion (Table 6).

Another finding relevant for answering the second research question refers to the discovery of a knowledge-accountability factor (Table 4). It is forged by knowledge getting, knowledge giving, information seeking, and accountability sharing. The link between these key informa-

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | IT usage | | | | | | | | | | | | | | | |
| 2 | Formalization | −.26** | | | | | | | | | | | | | | |
| 3 | Latent spatial disp. | .06 | −.18* | | | | | | | | | | | | | |
| 4 | Knowledge giving | .11 | −.13 | .03 | | | | | | | | | | | | |
| 5 | Role ambiguity | .10 | −.28** | .03 | −.08 | | | | | | | | | | | |
| 6 | Accountability sharing | .13 | −.08 | .11 | .18* | −.25** | | | | | | | | | | |
| 7 | Hierarchy | .01 | −.00 | .07 | −.05 | .20* | .08 | | | | | | | | | |
| 8 | Centralization | −.35** | −.00 | −.00 | .02 | −.11 | −.11 | −.09 | | | | | | | | |
| 9 | Trans-boundary com. | .27** | −.09 | .02 | −.04 | −.07 | .20* | .13 | −.30** | | | | | | | |
| 10 | Information seeking | −.13 | −.01 | .00 | .18* | −.05 | .09 | .02 | .00 | .00 | | | | | | |
| 11 | Knowledge getting | .02 | −.07 | .09 | .18* | .09 | .19 | .11 | −.12 | .18* | .33** | | | | | |
| 12 | Trust sharing | .20** | −.06 | −.25** | .13 | −.10 | .08 | .10 | −.05 | .20* | .10 | .08 | | | | |
| 13 | Expert power | .11 | −.02 | .04 | .14 | −.00 | .03 | .05 | .02 | .16 | −.14 | .02 | .08 | | | |
| 14 | Age | .25** | −.16 | −.00 | .09 | .07 | −.10 | −.09 | −.32** | −.05 | .09 | .27** | −.34** | −.06 | | |
| 15 | Firm experience | .28** | −.06 | .03 | .12 | −.00 | .10 | .01 | −.26** | .06 | .18* | .24 | .02 | −.08 | .60** | |
| 16 | Field experience | .20* | −.14 | −.01 | .11 | .07 | −.11 | .02 | −.26** | −.02 | .17 | .18* | −.01 | −.14 | .72** | .83** |
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |

* $p < 0.05$; ** $p < 0.01$ or better.

FIG. 4. Zero-order correlations.

Table 4. Variable sets resulting from factor analysis.*

| Variable | Factor: IT—communication | Knowledge—accountability | Role ambiguity | Demographics |
|---|---|---|---|---|
| Field experience | | | | 0.91 |
| Age | | | | 0.85 |
| Firm experience | | | | 0.84 |
| Trans-boundary communication | 0.71 | | | |
| IT usage | 0.65 | | | |
| Centralization | −0.50 | | | −0.44 |
| Trust sharing | 0.45 | | −0.45 | |
| Information seeking | | 0.67 | | |
| Knowledge getting | | 0.63 | | |
| Accountability sharing | | 0.56 | | |
| Knowledge giving | | 0.50 | | |
| Formalization | | | −0.65 | |
| Latent spatial dispersion | | | 0.62 | |
| Role ambiguity | | | 0.61 | |
| Eigenvalue (% of variance explained) | 1.83 (11.44) | 1.67 (10.41) | 1.51 (9.42) | 2.76 (17.24) |

The numbers in the table are factor loadings; loadings below 0.40 are omitted. The principal component method was used for determining the number of factors, which were rotated by the varimax method. The oblique method oblimin, available in SPSS version 7.5 for Windows, was also tried and gave the same solution. The four factors explain 50% of the variance, and addition of other factors adds much smaller increments to the explanation.

tion variables and team-based accountability, a dimension reflecting the position of work teams within the organizing and management domain, is interesting and important. Knowledge getting and accountability sharing appear to be key dimensions within this variable set. Knowledge getting builds the strongest relationship with information seeking, and acts as the second best predictor of accountability sharing. It also mediates (amplifies) relationships within the set, as well as outside it, acting as a suppresser to the relationship between trans-boundary communication and expert power, which otherwise is an isolated relationship (Table 5).

Accountability sharing, qualitative data indicate, is being practiced among team members in a broad host of critical issues, such as defining the scope of work, controlling team budget and organizing work. This last item can be effectively illustrated with the following statement describing an instance of grounding accountability in the work team context: "Each member's duties were delegated by the manager. However, after discussing what needed to be done, the team reassigned duties. The manager asked why. We explained and he agreed." Accountability sharing amplifies the relationship between knowledge getting and knowledge giving (Table 5). It also represents a bridge to the role ambiguity set, by forming a negative relationship with the variable of role ambiguity.

From the perspective of differentiating the non-traditional organization from its bureaucratic antipode, relevant are the relationships evolving around role ambiguity—the negative relationship with formalization, and the positive relationship with hierarchy (see Fig. 3). Both call into question fundamental presuppositions of the bureaucracy. Role ambiguity mediates relationships among dimensions that load on a factor that is named after this variable. However, partialling out role ambiguity does not make the correlation coefficients significant (Table 5).

A demographic factor is salient in the data set; centralization also loads on it, albeit weakly (Table 4). The typical respondent was male, between 20 and 29 years of age, with 4.5 years of experience in the field and 2.5 years spent with the current firm. The demographics variables are strongly inter-related, and all intervene in a number of relationships outside their set; both experience variables appear to be somewhat more influential than the age variable (Table 5). Specifically, all of the demographic variables suppress (to the level of insignificance) the relationship between expert power and trans-boundary communication. In addition, field experience makes insignificant the relationship between information seeking and firm experience.

While the above findings have been expected for the most part on the basis of the research model, certain expectations were not realized. Specifically, it was expected that IT usage would relate inversely to hierarchy, while no relationship between these was found; also, the isolated position of expert power was somewhat surprising.

## 5. Discussion

This section discusses findings that are most relevant for answering the two research questions that guided the present study: (1) What role does IT play in the non-traditional organization? (2) What are the other important

Table 5. Modifying effects.†

| Control for | Relationship | Partial $r$ | Zero-order $r$ |
|---|---|---|---|
| Centralization | IT usage-trans-boundary communication | 0.17 | 0.27** |
| | IT usage-age | 0.16 | 0.25** |
| | IT usage-field experience | 0.12 | 0.20* |
| | IT usage-firm experience | 0.21* | 0.28** |
| Trans-boundary communication | IT usage-trust sharing | 0.16 | 0.20* |
| | IT usage-centralization | −0.29** | −0.35** |
| IT usage | Centralization-firm experience | −0.18* | −0.26** |
| | Centralization-field experience | −0.20* | −0.26** |
| | Centralization-age | −0.24** | −0.32** |
| Age | IT usage-firm experience | 0.18* | 0.28** |
| | IT usage-field experience | 0.04 | 0.20* |
| | Centralization-firm experience | −0.10 | −0.26** |
| | Centralization-field experience | −0.06 | −0.26** |
| | Centralization-trans-boundary communication | −0.37** | −0.30** |
| Field experience | IT usage-firm experience | 0.21* | 0.28** |
| | IT usage-age | 0.15 | 0.25** |
| | Centralization-age | −0.18* | −0.32** |
| | Centralization-firm experience | −0.08 | −0.26** |
| | Age-firm experience | −0.01 | 0.60** |
| IT usage | Trans-boundary communication-trust sharing | 0.16 | 0.20* |
| | Formalization-latent | | |
| | Spatial dispersion | −0.16 | −0.18* |
| Knowledge getting | Knowledge giving-information seeking | 0.12 | 0.18* |
| | Expert power-trans-boundary communication | 0.24** | 0.16 |
| | Information seeking-firm experience | 0.12 | 0.18* |
| | Accountability sharing-knowledge giving | 0.15 | 0.18* |
| Centralization | Trans-boundary communication-expert power | 0.25** | −0.16 |
| | Trans-boundary communication-accountability sharing | 0.13 | 0.20* |
| | Knowledge getting-field experience | 0.12 | 0.18* |
| Age | Trans-boundary communication-expert power | 0.24** | 0.16 |
| | Accountability sharing-firm experience | 0.20* | 0.10 |
| | Information seeking-firm experience | 0.13 | 0.18* |
| Firm experience | Trans-boundary communication-expert power | 0.26** | 0.10 |
| | Accountability sharing-field experience | −0.28** | −0.11 |
| | Field experience-age | 0.49** | 0.79** |
| | Age-centralization | −0.22** | −0.32** |
| Field experience | Trans-boundary communication-expert power | 0.26** | 0.16 |
| | Accountability sharing-firm experience | 0.30** | 0.10 |
| | Centralization-age | −0.18* | −0.32** |
| | Information seeking-firm experience | 0.08 | 0.18* |
| Accountability sharing | Knowledge getting-knowledge giving | 0.13 | 0.18* |
| Trans-boundary communication | Accountability sharing-knowledge getting | 0.15 | 0.18* |
| Role ambiguity | Accountability | | |
| | Sharing-formalization | −0.17 | −0.08 |
| | Accountability | | |
| | Sharing-hierarchy | 0.14 | 0.08 |

† The modifying effects are determined by means of partial correlation analysis. $*p < 0.05$; $**p < 0.01$ or better.

characteristics of the non-traditional organization? Before beginning the discussion, several important limitations of the present study should be mentioned. Specifically, the sizes of samples taken are rather small, and the sample of local offices was geographically limited and not entirely random. In addition, the data collected are cross-sectional, which suits the exploratory character of the study but poses limitations to causal reasoning. Furthermore, due to the lack of appropriate constructs and measures, a significant number of them were newly developed, which in turn implies certain validity and reliability concerns. Finally, new organizational designs in general, and the non-traditional organization in particular, are complex research phenomena whose explanation requires perhaps more complex models than the one used here. This is apparent from the fact that both dyadic partial correlations and various regression models explain smaller amounts of the variance—at best, 10 and 27%, respectively.

### 5.1. The Nexus Role of Information Technology

The discovered high positive correlation between IT usage and the non-traditional organizational dimensions (partial $r = 0.68$) are in agreement with the literature that

Table 6. Predictors.*

| For the variable | Best predictor | Regression equation | $R^2$ |
|---|---|---|---|
| IT usage | Centralization | $y = 56.5 - 3.75 \times \text{Cen} - 1.8 \times \text{For} + 0.82 \times \text{Tru} + 0.73 \times \text{Yco}$ | 27 |
| Centralization | IT usage | $y = 6.66 - 0.004 \times \text{IT} - 0.19 \times \text{Com} - 0.52\ \text{Age}$ | 24 |
| Trans-boundary com. | IT usage | $y = 2.33 + 0.032 \times \text{IT}$ | 8 |
| Trust | Latent spatial dispersion | $y = 19.7 - 0.33 \times \text{Lsd} + 0.04 \times \text{IT}$ | 15 |
| Accountability | Role ambiguity | $y = 12.66 - 0.16 \times \text{Rol} + 0.33\ \text{Kge}$ | 10 |
| Knowledge getting | Information seeking | $y = 0.72 + 0.21 \times \text{Ise} + 0.58 \times \text{Age} + 0.11\ \text{Acc}$ | 18.3 |

* % of variance explained. Cen = centralization; For = formalization; Tru = trust sharing; Yco = firm experience; Com = trans-boundary communication; Lsd = latent spatial dispersion; Rol = role ambiguity; Kge = knowledge getting; Ise = information seeking; Acc = accountability sharing.

places IT at the nexus of new organizational designs. The organizations studied appear to be what Galbraith and Lawler (1993) call an "information-rich organization"—one that is capable of managing and taking advantage of the "new world of information complexity and richness" (pp. 296–297). Professionals in these organizations possess IT skills, deep understanding of the importance of information, and the competence of accessing, using and creating information—all critical potencies in Galbraith's characterization of the information-rich organization. In accord with Galbraith & Lawler's (1993) conceptualization, the distinctive IT role in the organizations studied instantiates in the relationships in which IT usage participates—two negative with centralization and formalization, and two positive with trust sharing and trans-boundary communication. All these variables but formalization load on the same larger dimension, termed "IT-communication" factor, because of the importance of the IT and the communication variables, whose explanatory power is the second largest (eigenvalue = 1.83). Each of the four relationships will be discussed in turn.

The inverse relationship between IT usage and centralization (concentration of control in superiors' hands) is essential because it deals with the precinct of power which is essential to any organizational design, including the new designs (partial $r = -0.29$; each variable is the best predictor of the other). This finding corroborates the evidence of the negative association between IT and centralization at the operational level in smaller organizations (e.g., Carter, 1984; Dawson & McLaughlin, 1986; Keon et al., 1992; Pfeffer & Leblebici, 1977). Where IT coincides with diminished management control at the operational level of production, changes in division of labor and authority are likely to be the antecedent—changes that Zuboff (1988) attributes to the advent of the "informated organization." In the case of the organizations studied, the professionals' conviction that a computer with a modem allows for "setting up an office anywhere" can be insightful. Even though this design has no official seal of approval, it is extant in intentions and actions. The respondents have pointed to the technological basis of this organizational design, and have provided an account of the computer being an indispensable box of computing and communication tools. Organizational and management implications of linking the computer with the new design cut to the heart of the present study: IT enables a new organization of work which brings the combined characteristics of the network, adhocracy, and organic organization to the fore. This design also affects distribution of control. It is pushed down to the operational level, to those in production that use production tools more intensely—hence the negative relationship between centralization and IT usage. One may argue that the basis of control is probably being shifted from legitimate power to some other bases. Heckscher (1994) proposes that this may be a "persuasive ability" which rests on equality of resources and knowledge/information that is relevant for dealing with a question at hand. The basis of this new control is an interesting question for future research, which was not investigated in the present study.

Another important finding for understanding the role of IT in the non-traditional organization is the inverse relationship between IT usage and formalization (the extent of written rules and regulations; partial $r = -0.22$). The literature provides inconclusive evidence and predictions regarding the relationship between IT and formalization; some authors view IT as a force that increases formalization (Huber, 1984; Wijnhoven & Wassenaar, 1990); others believe IT can decrease formalization (McKenney, 1985, as

Table 7. Fundamental technologies.

| Category | No. of respondents ($N = 131$) | Percentage of all respondents |
|---|---|---|
| Computer | 93 | 71 |
| Telephone/voice mail | 11 | 8 |
| Other | 21 | 16 |
| No answer | 6 | 5 |
| Total | 131 | 100 |
| **Reasons for importance of computers ($N = 93$)*** | | |
| Indispensable in general[a] | 34 | 28 |
| Necessary work tool | 32 | 27 |
| Necessary communication tool | 27 | 23 |
| Storage space | 16 | 13 |
| Efficiency vehicle | 11 | 9 |
| Total | 120 | 100 |

* The total number of respondents in the lower pane exceeds 93 because some answers were coded into more than one category.

[a] Examples: "My computer is the most important thing I need" or "All my work is in my computer."

cited in Wijnhoven & Wassenaar, 1990), while yet others establish no relationship between the two dimensions (Huber, 1990; Pfeffer & Leblebici, 1977). The present study can contribute to our understanding of this problem: Apparently, it corroborates McKenney's, 1985 assumption (as cited in Wijnhoven & Wassenaar, 1990, p. 51), and it stipulates that IT with both computing and communication capabilities can be associated with a lower self-reported formalization in small, team-based organizations. This relationship, then, becomes another characteristic of the non-traditional organizational design. In interpreting this relationship, it may be argued that the use of IT could result in making rules and regulations tacit. Heydebrand (1989) argues that formal rationality, calculability, and procedures can today be embedded in computer software, which, in turn, is conducive to increased self-regulation and more flexible forms of social control. In the organizations studied, the rules and regulations that drive the communication and data transfer tasks are, indeed, embedded in software; for example, shared database software can control document completion, accuracy, and changes across related documents; or, a highly usable user interface to a file transfer application can apply a host of rules of data protection, conversion, and authentication, while hiding this from the user. Although new rules have, therefore, been generated, these are not imposed on the worker-IT user in an overt fashion, which can consequently influence his perception of the degree of formalization. This capability of making rules tacit applies to domains other than communication as well, although these have not been the subject of the present study; for example, an expert system for conducting auditing procedures "hides" in itself a host of organizational and environmental rules and regulations. Another possible explanation has to do with the domain of action that the IT studied is capable of supporting, which is discussed later.

Both inverse relationships that IT forms with the structural dimensions characterizing a traditional organization—centralization and formalization—help us understand the structure of the non-traditional organization. Equally important for understanding this organization are the direct relationships IT builds with two cultural dimensions—the non-traditional dimension of trust sharing and trans-boundary communication which facilitates action and binds the non-traditional organization.

The direct relationship between IT usage and trust sharing (the expectation that co-workers' statements can be relied upon) adds another piece of evidence to the nexus role of IT in the non-traditional organization. The present study found that the significance of the relationship between IT usage and trust sharing is dependent on trans-boundary communication (if it is partialled out, the relationship between IT usage and trust sharing drops from 0.20, $p < .05$, to 0.16, $p < .07$). This dependence is consistent with the commonly shared thesis that communication underpins trust development and maintenance. IT usage, trust sharing and trans-boundary communication are, therefore, mutually related, and contribute to the non-traditional organizational design. The finding that trust is not likely to be diminished with increased IT usage is particularly important for the non-traditional organization which profoundly relies on IT. Trust has been considered a defining characteristic of new organizational designs, or the postmodern organization (Clegg, 1990), while its negation, mistrust, is compared with cholesterol that is clogging and incapacitating organizations (Blunt, 1989). In professional organizations such as those studied here, trust builds on understanding that the fortunes of all depend on combining the performances of all (Heckscher, 1994). This is typical of task culture, which is characterized by organizational members' focus on projects, collaboration inside the organization, and competitiveness outside of it (see Handy, 1993).

The link between trust sharing and IT usage can, therefore, be understood as a characteristic of a task culture with strong elements of collaboration. Moreover, the interdependence of tasks in information work, which is conducted in the organizations studied, implies a peculiarity, stemming from the need for a critical approach to data/information. Specifically, the information worker is in the position to trust the result of a colleague's work, so that she should (paradoxically) believe that her colleague has demonstrated a respectful lack of trust in the data/information the colleague has processed and passed on to her. The intermediation of IT in information flows between these two information workers potentially adds a layer of complexity to the maintenance of trust, because, at minimum, the intermediation reduces the possibility of immediate feedback and could, in itself, be a source of errors. This situation punctuates the importance of the discovered positive relationship between IT usage and trust. Caution is, however, recommended, since the present study did not investigate face-to-face communication, nor the proportion between it and the IT-mediated communication. It is plausible, however, that some mix of these two communication modes is most likely needed for supporting trust in new organizational designs (see Handy, 1996, pp. 211–214; Heckscher & Donnellon, 1994).

Also important is the finding of the positive association between IT usage and trans-boundary communication—the extent of team members' communicating beyond team boundaries (the measure of the association between IT usage and trans-boundary communication is the partial $r = 0.17$, and the significance of the relationship is dependent on centralization; also, IT usage is the best predictor of trans-boundary communication). This communication is carried over boundaries that separate an engagement team, the basis of work organization, from its office environment. IT provides one channel for this communication, which, in turn, has a significant role in the non-traditional design. The role of IT, therefore, refers to its support of this communication dimension. Since this dimension represents an important characteristic of the non-traditional organization, which answers partially the second research question, it will be discussed in the following section.

### 5.2. *Communication as Panopticon, Spinal Cord, and Dialog*

The multiple relationships that trans-boundary communication participates in (five in total) make it appear a panopticon, in which diverse layers of organizational reality reflect themselves—IT, knowledge acquisition (knowledge getting), structure (centralization), politics (expert power), behavioral patterns related to the management of work (accountability sharing), and the psycho-social element of organizational culture (trust sharing). These are key dimensions in the non-traditional organization that have relevance for other new designs as well. But this communication plays more than a mirroring role.

The metaphor of the spinal cord is being proposed here in order to convey how instrumental this communication is in binding the non-traditional organization together. Propositions regarding the network organization presaged by Rockart and Short (1991), Morgan (1993), Galbraith & Lawler (1993) and others are manifest here: It indeed is communication that binds work teams with the rest of the organization. As the findings of the present study suggest, this communication is used for knowledge acquisition and in accountability matters. It helps establish the Archimedes point between the Scylla of groupthink and the Charybdis of underdeveloped team identity. Owing to this communication, organizational units can be "completely connected while remaining far apart" (Morgan, 1993, p. 10), and sharing of knowledge (Quinn, 1992), vision and values (Morgan, 1993), goals, decision making, accountability, and trust (Rockart & Short, 1991) are all being facilitated through this communication. The role spinal cord-like communication plays in new designs can also be interpreted in terms of action that affects organizational structure. This self-governed communication coalesces with self-organizing practices and constitutes a domain of action, or communicative action (Habermas, 1984, 1987). Communicative action is not only different from structure-driven behavior, but impinges on the established structure, as has been demonstrated in the organizations studied. This is where the adhocratic character of organization also surfaces, because action is driven by the purposes at hand, and so work roles are being reshaped dynamically according to these purposes.

Metaphors of dialogue and text could also inform our understanding of the place of trans-boundary communication in relation to knowledge acquisition, IT usage, and other dimensions in the non-traditional organization. For Heckscher (1994), dialogue is a cornerstone of an "interactive organization," a vehicle for developing principles of action that replaces traditional rules and regulations, and underpins the sharing of trust, which, in turn, is the basis of the persuasive ability substituting for power relationships and legitimizing the necessary hierarchy. There is an apparent symmetry between these propositions and the relationships discovered in the present study. The professionals studied engage in communication beyond team boundaries when task-related needs demand (indicated by the positive relationship between trans-boundary communication and trust sharing), avert power relationships (indicated by the isolate position of expert power and its suppression by knowledge getting), and legitimize a necessary hierarchy (indicated by the positive relationship between trans-boundary communication and accountability sharing—part of this communication may be directed toward superiors when the need of justifying team results arises).

The juxtaposition of dialogue and text that Taylor and Van Every (1993) discuss can also inform our understanding of the role of trans-boundary and other communication aspects in the organizations studied. While text induces rather an inflexible, rules-driven context, dialogue resembles an ongoing, live, action-rich organizational context. Dialogue is thus more like free-flowing communication, which helps organization members make sense of the organizational whole and seek mutually agreeable co-alignments. Extending this metaphorical reasoning into the information perspective, dialogue can be viewed as a generator of information. Dialogue manifests itself through an organizational sphere, in which any meaning is created intersubjectively, boundaries around work problems are consensually drawn, and pipes, filters, and switches (to borrow Unix terminology) for interpretation of data are negotiated. Dialogue increases the scope, complexity, and unpredictability of information. Text, on the other hand, reduces unpredictability by constraining information options; also, it generates costs in the form of petrifying organizational structure and depressing intellective interactions to a low bandwidth.

In the organizations studied, text is represented, for example, in standard, proprietary work procedures, computer software, IT know-how, standards, relevant legislature (e.g., that which determines the scope of auditing), and business plans. This text constitutes what may be called the infrastructure of the organization. When a work team is delegated a project, a dialogue between the team and the superior can ensue in order to define detailed work organization and the timeline (as is demonstrated in the example of negotiating accountability, which has been discussed). The team members also conduct an internal dialogue among themselves in order to define individual roles—the structure that will overlay the infrastructure; this structure can contain flexible and changeable roles, which make it adhocratic in character. This internal dialogue continues throughout the project work, and instantiates, among other activities, in knowledge exchanges and information seeking. The same information activities motivate team members to engage in a dialogue with outsiders. Dialogue, therefore, develops on the ground of the self-defined structure, which is also partly created through dialogue, and which rests on the infrastructure or text. Understood in this way, dialogue invokes, again, Habermas's concept of communicative action (1984, 1987). Text, on the other hand, appears to be rather buffered from ongoing work, and serves as an underlying carrier for perpetual dialogue that coalesces with work. Text comes

first, but dialogue is indispensable for organizing around projects and completing them. Dialogue therefore represents another characteristic of a task-collaborative culture.

### *5.3. Team Context: Accountability, Knowledge, and Roles*

Another important characteristic of the non-traditional organization refers to accountability sharing (the extent of a team member's readiness to justify the team's decisions to superiors). The present study has found that this accountability is in relation with knowledge giving and getting, and information seeking (all these load on the same factor). Trans-boundary communication is partially involved in these relationships (the significance of the relationship between accountability sharing and knowledge getting is dependent on trans-boundary communication). The discovered link between team accountability and rich professional knowledge is in agreement with the literature of new organizational designs. This relationship is expected, because group accountability is only possible where multiple skills and flexible skills rather than restrictive skill defensiveness are the norm. In a context with a high skill division, which is in stark contrast to the organizations studied, individual roles are emphasized, which, in turn, calls for more formalized and externalized coordination and control (see Clegg, 1990).

Teams in the organizations studied are held accountable for a host of important management issues, including defining the scope of work, control of the team budget, and organization of work. The scope and content of team accountability demonstrates the capacity and the depth of the self-organizing in these organizations. This broad team accountability also demonstrates the magnitude of adhocracy characteristics intrinsic to this organization: Although a team is held accountable by superiors for the whole project, individual roles within the team have to be designed and redesigned as work needs at hand demand. Moreover, the fact that the variables of knowledge exchange, information seeking, and accountability sharing form a separate set suggests that information activities are important for the teams' responding to accountability obligations. Exchanges of knowledge among equals may also be instrumental in explaining why expert power is suppressed by knowledge getting and the demographics factor. Specifically, it may well be that the older and more experienced respondents do not feel that expert power is at work in their knowledge-getting activities, because a broader distribution of knowledge through its self-regulated exchanges reduces the basis for power differentiation in organizations (cf. Zuboff, 1988). The accountability dimension and its relationships explain, furthermore, a task-collaboration culture that suits the non-traditional organization.

Finally, it is interesting and important that role ambiguity (the lack of information necessary to an organizational position) relates negatively both to hierarchy (the extent of graded authority levels) and formalization. These findings add to the explanation of the non-traditional structure, which was discussed in the section on IT above. The professionals in the organizations studied, therefore, believe that more management levels and written rules do not contribute to the clarification of work-related expectations attached to the professionals. Where hierarchy and formalization create ambiguity of work roles, the ontology of the bureaucratic organization suffers another serious fracture (Taylor & Van Every, 1993). In the Weber ideal model of bureaucracy, the reason d' être of layered management and written rules is precisely the opposite—to create and manage clear division of labor. In the organizations studied, however, work teams are, for the most part, in charge of shaping work roles—there is a larger expectation which a superior attributes to a whole team, while the specific definitions of individual roles are up to the team. With regard to rules and regulations that apply to a particular domain of information work, these are either internalized through practice or built into computer software, which diminishes the perception of their external existence. More specific implications of the inverse relating of role ambiguity to hierarchy and to formalization for information production and management, however, need to be addressed in future research.

### *5.4. Hierarchy Resilience and Propinquity Limitations*

Although the organizations studied corroborate a larger part of the model of the non-traditional organization used, a certain deviation from the model has been identified. Most apparently, the average hierarchy in the organizations studied is four, even though these organizations are small. This hierarchy deviates both from the current trends of flattening organizational structure and from the arguments on the cessation of the status quo (Clemons & Row, 1989), on the rise of equivocality in the organizational environment (Castels, 1986; Weick, 1979), and on a shift in management philosophies from running hierarchies to running more horizontal organizations of people who learn and who are creative (Taylor & Van Every, 1993). Moreover, the state of hierarchy found matches certain historical instances of hierarchy in traditionally more hierarchical organizations, such as the government organizations investigated by Blau and Schoenherr (1971) and small factories studied by Pfeffer and Leblebici (1977).

To clarify, an "infinitely flat organization" (Quinn, 1992) is perhaps not a solution for the organization in which professional experience and work quality demand a certain vertical differentiation. This is what Heckscher (1994) would call a "necessary hierarchy," whose extent could be determined by the technical criterion of the hierarchy's contribution to resolving production-related uncertainty (see Galbraith, 1973). One of these regulatory criteria could be the amount of IT usage at the operational level of information work, and so a negative association between IT and hierarchy may be anticipated, as the literature suggests (Drucker, 1987; Huber, 1990; Pool, 1983; Whisler, 1970; Wigand, 1985). However, hierarchy appears to be resilient

in this respect too, since no relationship between the two dimensions was discovered. It might be that technical criteria mix with ownership concerns in the public accounting industry. Since upper management ranks are blended with the partnership-related rights and responsibilities, it could be that this type of ownership is an additional source of maintaining a more traditional hierarchy. This issue certainly demands more research. At any rate, the coexistence of hierarchy and a number of the non-bureaucratic characteristics discussed here is likely to create tensions that have been reported in the literature of organization of the accounting industry (Pratt & Jiambalvo, 1981; Sorensen, 1967; Sorensen & Sorensen, 1974; Watson, 1975). The resilience of hierarchy, therefore, could be understood as a peculiarity of the sample/population studied. However, Heckscher and Donnellon (1994) argue that the same phenomenon can be identified in a broader organizational world. If true, then the finding of the present study can be understood so that the extent of hierarchy comprises one of the key indicators for testing organizations on the non-traditional design. This should be addressed in further research.

Another impediment to the non-traditional organization is signified by the negative relationship between trust sharing and latent spatial dispersion (the self-reported propensity of working outside the office space). It follows that the mediation of IT in communication, which underpins trust maintenance, meets spatial limits. This can have repercussions for the design of setting up an office anywhere, which some respondents mentioned as their preferred sort of organization. One should keep in mind that the face-to-face context is predominant in the organizations studied, and so it is unclear how trust and spatial dispersion would relate in the new designs more accustomed to a distributed mode, such as virtual organizations. In any case, the spatial dimension of organizing will certainly remain a critical issue in any new organizational design. Krackhardt's (1994) invocation of the law of propinquity (the frequency of communication among organization members decreases more than linearly with distance) could sound even fatalistic in this regard. He speculates that this law could establish unwavering limitations for new organizational designs. But it is equally reasonable to assume that the idea of setting up an office anywhere need not necessarily go awry. To become materialized, assiduous design and maintenance are necessary, including the provision for face-to-face communication and technological conditions for synchronizing time frames and work flows. This should certainly concern both general and information systems management.

## 6. Conclusion

IT appears to be at the nexus of the non-traditional organization. IT is related inversely to two traditional structural variables—centralization and formalization. IT enables a new organization of work, which some respondents described as the possibility of "setting up an office anywhere." This organization brings the combined characteristics of the organic, network, and organic organization to the fore, and pushes control down to the operational level. The inverse relationship between IT usage and normalization could be because IT can make organizational rules and regulations tacit, built into software, and, so, transparent to the user. Another part of the explanation of the non-traditional structure refers to the inverse relationships role ambiguity builds with hierarchy and with formalization. Where hierarchy and formalization create ambiguity of work roles, the ontology of the bureaucratic organization suffers another serious fracture. Moreover, IT usage is in direct relationship with trust sharing and trans-boundary communication. The first relationship is particularly important for this IT-dependent organization, in which one should trust results of a colleague's information work, even though questioning the data/information is part of the job. On the other hand, IT supports communication beyond team boundaries, resembling a spinal cord; this communication is being used for knowledge acquisition and in responding to team accountability obligations. In addition, this communication resembles a dialogue through which team roles are negotiated with superiors, and knowledge and information are exchanged. It also can be understood as communicative action that impinges on and shapes organizational structure. The culture of the non-traditional organization can be understood as a task-collaboration culture, an essential characteristic of which is team-based accountability. This accountability applies to a host of key production activities, and is related to self-organized exchanges of knowledge and information.

Although these findings are, for the most part, expected, and fit the research model developed and used in the study, two important findings deviate from the framework of new designs: The hierarchy discovered appears to be at a discrepancy, both with the small organizational size and technical criteria. In addition, the dimension of spatial dispersion is negatively associated with trust sharing, which might signal spatial limitations to the idea of setting up an office anywhere. Further research needs to address the basis of control, the specific information implications of the negative associations that role ambiguity forms with hierarchy and with formalization, the connection between the state of hierarchy and the partnership type of ownership, the value of hierarchy as an indication of the non-traditional organizational structure, and the limitations to the spatial dimensions of organizing. In addition, IT needs to be investigated beyond its aspect of the amount of usage; more complex organizational models, with a greater explanatory power, ought to be used. All these can help broaden the initial insights into the information and concomitant aspects of the non-traditional organization, one of the new organizational designs whose time has come.

## Acknowledgment

My sincere thanks are extended to the colleagues who helped me through the present study—Jeffrey Katzer, Rolf Wigand, Murali Venkatesh, and Robert Benjamin, as well as to anonymous reviewers.

## References


Adams, D. A., Nelson, R. R., & Todd, P. A. (1992, June). Perceived usefulness, ease of use, and usage of information technology. *MIS Quarterly,* pp. 227–247.

Aldrich, H. (1981). Organizations, action sets, and networks: Making the most of simplicity. In P. C. Nostrom & W. H. Starbuck (Eds.), *Handbook of organizational design: Vol. 1.* New York: Oxford University Press.

Ashkenas, R., Urlich, D., Jick, T., & Kerr, S. (1995). *The boundaryless organization: Breaking the chains of organizational structure.* San Francisco: Jossey-Bass.

Bachman, J. G., Bowers, D. G., & Marcus, P. M. (1968). Bases of supervisory power: A comparative study in five organizational settings. In A. Tannenbaum (Ed.), *Control in organizations.* New York: McGraw-Hill.

Bailey, A., & Neilsen, E. H. (1992). Creating a bureau-adhocracy: Integrating standardized and innovative services in a professional work group. *Human Relations, 45*(7), 687–710.

Baker, W. E. (1992). The network organization in theory and practice. In N. Nohria & R. E. Eccles (Eds.), *Networks and organizations: Structure, form, and action* (pp. 397–429). Boston: Harvard Business School Press.

Ballew, V. (1982). Technological routines and intra-unit structure in CPA firms. *The Accounting Review, 57*(1), 88–104.

Barley, S. R. (1990). The alignment of technology and structure through roles and networks. *Administrative Science Quarterly, 35,* 61–103.

Benjamin, R., & Scott Morton, M. (1992). Reflections of effective application of information technology. In F. H. Vogt (Ed.), *Personal computers and intelligent systems. Information Processing 92,* III. North Holland: Elsevier.

Bennis, W. G., & Slater, P. E. (1968). *The temporary society.* New York: Harper & Row.

Blau, P., & Schoenherr, R. A. (1971). *The structure of organizations.* New York: Basic Books.

Blunt, P. (1989). In S. Clegg (Ed.), *Modern organizations: Organization studies in the postmodern world* (p. 202). London: Sage.

Burkhardt, M. E., & Brass, D. J. (1990). Changing patterns or patterns of change: The effects of a change in technology on social network structure and power. *Administrative Science Quarterly, 35,* 104–127.

Burns, T., & Stalker, G. M. (1961). *The management of innovation.* London: Tavistock.

Carter, N. M. (1984). Computerization as a predominate technology: Its influence on the structure of newspaper organizations. *Academy of Management Journal, 27*(2), 247–270.

Castels, M. (1986). High technology, world development, and structural transformation: The trends and the debate. *Alternatives, XI,* 297–343.

Chandler, A. D. (1962). *Strategy and structure.* Cambridge, MA: The MIT Press.

Child, J. (1987, Fall). Information technology, organization, and the response to strategic challenges. *California Management Review, Fall,* pp. 33–50.

Ciborra, C. U. (Ed.). (1996). *Groupware & teamwork: Invisible aid or technical hindrance?* Chichester, UK: Wiley.

Clegg, S. R. (1990). *Modern organizations: Organization studies in the postmodern world.* London: Sage.

Clegg, S. R. (1996). Introduction. In S. Clegg, C. Hardy, & W. R. Nord (Eds.), *Handbook of organizational studies.* Thousand Oaks, CA: Sage.

Clemons, E. K., & Row, M. C. (1989). Information technology and economic reorganization. Warton School of Business, University of Pennsylvania, Philadelphia. In J. R. Taylor & E. J. Van Every (Eds.), *The vulnerable fortress: Bureaucratic organizations and management in the information age* (p. 217). Toronto: University of Toronto Press.

Cushman, D. P., & King, S. S. (1995). *Communication and high-speed management.* Albany, NY: SUNY Press.

Dahl, R. A. (1957). *The concept of power.* Behavioral Science, 2, 201–215.

Davis, F. D. (1989, September). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly,* 319–340.

Dawson, P., & McLaughlin, I. (1986). Computer technology and the redefinition of supervision: A study of the effects of computerization on railway freight supervisors. *Journal of Management Studies, 23,* 116–132.

Dhar, V., & Olson, M. H. (1989). Assumptions underlying systems that support work group collaboration. In M. Olson (Ed.), *Technological support for work group collaboration.* Hillsdale, NJ: Erlbaum.

Dirsmith, M., & Covalevski, M. (1985). Informal communications, nonformal communications, and mentoring in public accounting firms. *Accounting, Organizations and Society,* pp. 149–169.

Donnellon, A., & Scully, M. (1994). Teams, performance, and rewards: Will the post-bureaucratic organization be a post-meritocratic organization? In C. Heckscher & A. Donnellon (Eds.), *The post-bureaucratic organization: New perspectives on organizational change* (pp. 63–90). Thousand Oaks, CA: Sage.

Drucker, P. F. (1987/1990). The coming of the new organization. In P. Drucker (Ed.), *The new realities in government and politics, in economics, in society and world.* New York: Harper & Row.

Eccles, R. G., & Crane, D. B. (1987, Fall). Managing through networks in investment banking. *California Management Review, Fall,* pp. 176–195.

Foulger, D. A. (1990). *Medium as process: The structure, use and practice of computer conferencing on IBM's IBMPC computer conferencing facility.* Unpublished doctoral dissertation, Temple University, Philadelphia, PA.

French, J., Jr., & Raven, B. (1959). The basis of social power. In D. Cartwright (Ed.), *Studies in social power* (pp. 150–167). Ann Arbor: University of Michigan.

Fry, L. W. (1982). Technology-structure research: Three critical issues. *Academy of Management Journal, 25*(3), 532–552.

Fulk, J., & DeSanctis, G. (1995). Electronic communication and changing organizational forms. *Organization Science, 6*(4), 337–349.

Galbraith, J. R. (1973). *Designing complex organizations.* Reading, MA: Addison-Wesley.

Galbraith, J. R., & Lawler, E. E., III. (1993). Effective organizations: Using new logic of organizing. In J. R. Galbraith & E. E. Lawler III. (Eds.), *Organizing for the future: The new logic of managing complex organizations* (pp. 285–299). San Francisco: Jossey-Bass.

Goldman, S. L., Nagel, R. N., & Preiss, K. (1995). *Agile competitors and virtual organizations: Strategies for enriching the customer.* New York: Van Nostrand Reinhold.

Guttiker, U. E., & Gutek, B. A. (1988). Office technology and employee attitudes. *Social Science Computer Review, 6*(3), 327–340.

Habermas, J. (1984). *The theory of communicative action, Vol. 1.* Boston: Beacon Press.

Habermas, J. (1987). *The theory of communicative action, Vol. 2.* Boston: Beacon Press.

Hammel, G., & Prahalad, C. K. (1994). *Competing for the future.* Boston, MA: Harvard Business School Press.

Hammer, M. (1996). *Beyond reengineering: How the process-centered organization is changing our work and our lives.* New York: Harper Business.

Hammer, M., & Champy, J. (1993). *Reengineering the corporation: A manifesto for business revolution.* New York: Harper Business.

Handy, C. (1989). *The age of unreason.* Boston: Harvard Business School Press.

Handy, C. (1993). *Understanding organizations.* New York–Oxford: Oxford University Press.

Handy, C. (1996). *Beyond certainty: The changing world of organizations.* Boston: Harvard Business School Press.

Heckscher, C. (1994). *Defining the post-bureaucratic type.* In C. Heckscher & A. Donnellon (Eds.), *The post-bureaucratic organization: New perspectives on organizational change* (pp. 14–62). Thousand Oaks, CA: Sage.

Heckscher, C., & Donnellon, A. (Eds.), (1994). *The post-bureaucratic organization: New perspectives on organizational change.* Thousand Oaks, CA: Sage.

Heydebrand, W. V. (1989). New organizational forms. *Work and occupations, 16*(3), 323–357.

Hiltz, S. R. (1984). *Online communities: A case study of the office of the future.* Norwood, NJ: Ablex.

Huber, G. P. (1984, August). The nature and design of post-industrial organizations. *Management Science, 30*(8), 928–951.

Huber, G. P. (1990). A theory of the effects of advanced information technologies on organizational design, intelligence, and decision making. *Academy of Management Review, 15*(1), 47–71.

Jaques, E. (1989). *The requisite organization: The CEO's guide to creative structure and leadership.* Kingston, NY: Cason Hall.

Kahn, R. L., Wolfe, D. M., Quinn, R. P., Snoek, J. D., & Rosenthal, R. A. (1964). *Organizational stress.* New York: Wiley.

Keon, T. L., Vazzana, G. S., & Slocombe, T. E. (1992). Sophisticated information processing technology: Its relationship with an organization's environment, structure and culture. *Information Resources Management Journal, 5*(4), 23–31.

Kerr, E. B., & Hiltz, S. R. (1982). *Computer-mediated communication systems: Status and evaluation.* New York: Academic Press.

Kiesler, S., Siegel, J., & McGuire, T. W. (1984). Social psychological aspects of computer-mediated communication. *American Psychologist, 39*(10), 1123–1134.

Kling, R. (1997). Available: http://www.slis.indiana.edu/CSI.

Krackhardt, D. (1994). Constraints on the interactive organization as an ideal type. In C. Heckscher & A. Donnellon (Eds.), *The post-bureaucratic organization: New perspectives on organizational change* (pp. 211–222). Thousand Oaks, CA: Sage.

Lash, S. (1988). Postmodernism as a regime of signification. *Theory, culture and society, 5*(2–3), 311–336.

Lawrence, P., & Lorsch, J. W. (1968). *Organization and environment: Managing differentiation and integration.* Boston: Harvard Business School.

MacDonald, A. P., Jr., Kessel, V. S., & Fuller, J. B. (1972). Self-disclosure and two kinds of trust. *Psychological Reports, 30,* 143–148.

Mankin, D., Cohen, S. G., & Bikson, T. K. (1996). *Teams and technology: Fulfilling the promise of the new organization.* Boston: Harvard Business Press.

Markham, W. T., Bonjean, C. M., & Corder, J. (1984). Measuring organizational control: The reliability and validity of the control graph approach. *Human Relations, 37,* 263–294.

Markus, L. (1994). Finding a happy medium. *ACM Transactions of Information Systems, 12*(2), 119–149.

Mintzberg, H. (1979). *The structuring of organizations: A synthesis of the research.* Englewood Cliffs, NJ: Prentice-Hall.

Mintzberg, H. (1983). *Power in and around organizations.* Englewood Cliffs, NJ: Prentice-Hall.

Mintzberg, H., & McHugh, A. (1985). A strategy formation in an adhocracy. *Administrative Science Quarterly, 30,* 160–197.

Morgan, G. (1993). *Imaginization: The art of creative management.* Newbury Park, CA: Sage.

Myles, R. E., & Snow, C. C. (1986). Organizations: New concepts for new forms. *California Management Review, 28*(3), 62–73.

Nohria, N., & Berkley, J. D. (1994). The virtual organization: Bureaucracy, technology, and implosion of control. In C. Heckscher & A. Donnellon (Eds.), *The post-bureaucratic organization: New perspectives on organizational change* (pp. 108–128). Thousand Oaks, CA: Sage.

Oldham, G. R., & Hackman, J. R. (1981). Relationships between organizational structure and employee reactions: Comparing alternative frameworks. *Administrative Science Quarterly, 26,* 66–83.

Olson, M. H., & Bly, S. A. (1991). The Portland experience: A report on a distributed research group. *International Journal of Man-Machine Studies, 34,* 211–228.

Peters, T. (1987). *Thriving on chaos.* New York: Knopf.

Peters, T. (1992). *Liberation management: Necessary disorganization for the nanosecond nineties.* New York: Knopf.

Peters, T., & Waterman, R. H., Jr. (1982). *In search of excellence: Lessons from America's best-run companies.* New York: Harper & Row.

Pfeffer, J. (1981). *Power in organizations.* Marshfield, MA: Pitman.

Pfeffer, J., & Leblebici, H. (1977). Information technology and organizational structure. *Pacific Sociological Review, 20*(2), 241–260.

Pool, de Sola, I. (1983). *Forecasting the telephone: A retrospective technology assessment.* Norwood, NJ: Ablex.

Powell, W. W. (1990). Neither market nor hierarchy: Network forms of organization. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (pp. 295–336). Greenwich, CT: JAI Press.

Pratt, J., & Jiambalvo, J. (1981). Relationships between leader behaviors and audit team performance. *Accounting, Organizations and Society, 6*(2), 133–142.

Price, J. L., & Mueller, C. W. (1986). *Handbook of organizational measurement.* Marshfield, MA: Pitman.

Quinn, J. B. (1992). *Intelligent enterprise: A knowledge and service based paradigm for industry.* New York: The Free Press.

Rice, R. E. (Ed.). (1984). *The new media.* Beverly Hills, CA: Sage.

Rice, R. E., & Love, G. (1987). Electronic emotion: Socio-emotional content in a computer-mediated communication network. *Communication Research, 12*(1), 85–88.

Rizzo, J. R., House, R. J., & Lirtzman, S. I. (1970). Role conflict and ambiguity in complex organizations. *Administrative Science Quarterly, 15,* 150–163.

Rockart, J. F., & Short, J. E. (1991). The networked organization and the management of interdependence. In S. S. Morton (Ed.), *The corporation of the 1990s: Information technology and organizational transformation.* Oxford: Oxford University Press.

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality, 35,* 651–665.

Shapiro, N. Z., & Anderson, R. H. (1985). *Toward an ethics and etiquette for electronic mail.* Washington, DC: Rand.

Siegel, J., Dubrovsky, V., Kiesler, S., & McGuire, T. W. (1986). Group processes in computer-mediated communication. *Organizational Behavior and Human Decision Processes, 37,* 157–187.

Smilowitz, M., Compton, D. C., & Flint, L. (1988). The effects of computer mediated communication on an individual's judgment: A study based on the methods of Asch's social influence experiment. *Computers in Human Behavior, 4,* 311–321.

Smolensky, M. W., Carmody, M. A., & Halcomb, C. G. (1990). The influence of task type, group structure and extroversion on uninhibited speech in computer-mediated communication. *Computers in Human Behavior, 16,* 261–272.

Sorensen, J. E. (1967, July). Professional and bureaucratic organization in the public accounting firm. *The Accounting Review,* July, pp. 553–566.

Sorensen, J. E., & Sorensen, T. L. (1974, March). The conflict of professionals in bureaucratic organizations. *Administrative Science Quarterly,* March, pp. 98–106.

Sproul, L., & Kiesler, S. (1991). *Connections: New ways of working in the networked organization.* Cambridge, MA: The MIT Press.

Stevens, M. (1981). *The big eight.* New York: Macmillan.

Stevens, M. (1985). *The accounting wars.* New York: Macmillan.

Sutton, S. G. (1992). Can we research a field we cannot define. In S. G. Sutton (Ed.), *Advances in accounting information systems: Vol. 1* (pp. 1–13). Greenwich, CT: JAI Press.

Tanenbaum, A. (1961). Control and effectiveness in a voluntary organization. *American Journal of Sociology, 67,* 33–46.

Taylor, J., & Van Every, E. (1993). *The vulnerable fortress: Bureaucratic organizations and management in the information age.* Toronto: University of Toronto Press.

Tetlock, P. E. (1983). Accountability and complexity of thought. *Journal of Personality and Social Psychology, 45*(1), 74–83.

Toffler, A. (1980). *The third wave.* New York: William Morrow.

Tushman, M. L., & Anderson, P. (1986). Technological discontinuities and organizational environments. *Administrative Science Quarterly, 31,* 439–465.

Watson, D. J. H. (1975, April). The structure of project teams facing differentiated environments: An exploratory study in public accounting firms. *The Accounting Review,* 259–273.

Weber, M. (1946). *Essays in sociology.* Oxford: Oxford University Press.

Weick, K. (1979). *The psychology of organizing.* New York: Random House.

Weinstein, G. W. (1987). *The bottom line: Inside accounting today.* New York: Nal Books.

Whisler, T. L. (1970). *The impact of computers on organizations.* New York: Praeger.

Wigan, M. (1991). Computer-cooperative work: Communications, computers and their contribution to working in groups. In R. Clarke & J. Cameron (Eds.), *Managing information technology's organizational impact* (pp. 15–24). Amsterdam: North-Holland.

Wigand, R. T. (1985). *Integrated communications and work efficiency: Impacts on organizational structure and power.* Paper presented at Beyond Polemics: Paradigm Dialogues, International Communication Association 35th Annual Conference, May 23–27, 1985, Honolulu, HI.

Wijnhoven, A. B. J. M., & Wassenaar, D. A. (1990). Impact of information technology on organizations: The state of the art. *International Journal of Information Management, 10,* 35–53.

Winograd, T. (1988). A language/action perspective on the design of cooperative work. In M. H. Olson (Ed.), *Technological support for work group collaboration.* Hillsdale, NJ: Erlbaum.

Zuboff, S. (1988). *In the age of the smart machine: The future of work and power.* New York: Basic Books.

# Call for Papers

## 1999 Information Resources Management Association International Conference

**May 16–19, 1999**
**Hershey, PA**

Developments in information technology have had a major impact on all types of libraries over the past two decades. The challenges for libraries that arise from the advent of the computer and the proliferation of cheaper, faster, more efficient means of accessing and processing data are particularly wide-ranging. Public, academic, and special libraries are faced with dilemmas as tightly focused as whether to opt for online or CD-Rom technology to complex ethical questions about the public's right to know. This IRMA track seeks to expand knowledge in this rapidly evolving area. Library practitioners and researchers are encouraged to submit papers.

*Some topics include:*

- Electronic resources sharing
- Digital libraries
- Virtual libraries
- Access versus possession of information
- The role of the librarian in a digital age
- Technology for special libraries
- Electronic publishing
- Full-text document and citation searches
- Security issues for libraries on the Internet
- Natural language processing
- The right to know—public access to information

*Submission Categories* include: Full length submissions of previously unpublished, conceptual or empirical research manuscripts; research-in-progress submissions; and panel, workshop, tutorial, and symposium submissions.

*Submission Guidelines* can be found at: *http://www.irma-international.org*

**Dates:**

| | |
|---|---|
| Deadline for receipt of Papers/ Proposals: | October 12, 1998 |
| Notification of acceptance/rejection: | November 29, 1998 |
| Deadline for receipt of final papers: | January 18, 1999 |
| Early registration ends: | April 3, 1999 |

**All inquiries about this track are to be made to:**

Patricia Diamond Fletcher, Ph.D. Track Chair
The University of Maryland, Baltimore County
1000 Hilltop Circle
Baltimore, MD 21250
(410) 455-3154 (office)
(410) 455-1073 (fax)
fletcher@umbc.edu

**All submissions are to be sent to:**

Mehdi Khrosrowpour, Program Chair
1999 IRMA International Conference
1331 E. Chocolate Avenue
Hershey, PA 17033-1117
(717) 533-8879 (office)
(717) 533-8661 (fax)
mehdi@irma-international.org

DISK SUBMISSION INSTRUCTIONS

## Please return your final, revised manuscript on disk as well as hard copy. The hard copy must match the disk.

The Journal strongly encourages authors to deliver the final, revised version of their accepted manuscripts (text, tables, and, if possible, illustrations) on disk. Given the near-universal use of computer word-processing for manuscript preparation, we anticipate that providing a disk will be convenient for you, and it carries the added advantages of maintaining the integrity of your keystrokes and expediting typesetting. Please return the disk submission slip below with your manuscript and labeled disk(s).

**Guidelines for Electronic Submission** (also available at http://www.interscience.wiley.com)

*Text*
**Storage medium.** 3-1/2″ high-density disk in IBM MS-DOS, Windows, or Macintosh format.

**Software and format.** Microsoft Word 6.0 is preferred, although manuscripts prepared with any other microcomputer word processor are acceptable. Refrain from complex formatting; the Publisher will style your manuscript according to the Journal design specifications. Do not use desktop publishing software such as Aldus PageMaker or Quark XPress. If you prepared your manuscript with one of these programs, export the text to a word processing format. Please make sure your word processing program's "fast save" feature is turned off. Please do not deliver files that contain hidden text: for example, do not use your word processor's automated features to create footnotes or reference lists.

**File names.** Submit the text and tables of each manuscript as a single file. Name each file with your last name (up to eight letters). Text files should be given the three-letter extension that identifies the file format. Macintosh users should maintain the MS-DOS "eight dot three" file-naming convention.

**Labels.** Label all disks with your name, the file name, and the word processing program and version used.

*Illustrations*
All print reproduction requires files for full color images to be in a CMYK color space. If possible, ICC or ColorSync profiles of your output device should accompany all digital image submissions.

**Storage medium.** Submit as separate files from text files, on separate disks or cartridges. If feasible, full color files should be submitted on separate disks from other image files. 3 1/2″ high-density disks, CD, Iomega Zip, and 5 1/4″ 44- or 88-MB SyQuest cartridges can be submitted. At authors' request, cartridges and disk will be returned after publication.

**Software and format.** All illustration files should be in TIFF or EPS (with preview) formats. Do not submit native application formats.

**Resolution.** Journal quality reproduction will require greyscale and color files at resolutions yielding approximately 300 ppi. Bitmapped line art should be submitted at resolutions yielding 600–1200 ppi. These resolutions refer to the output size of the file; if you anticipate that your images will be enlarged or reduced, resolutions should be adjusted accordingly.

**File names.** Illustration files should be given the 2- or 3-letter extension that identifies the file format used (i.e., .tif, .eps).

**Labels.** Label all disks and cartridges with your name, the file names, formats, and compression schemes (if any) used. Hard copy output must accompany all files.

**Detach and return with labeled disk(s)**

---

Corresponding author's name: ____________ E-mail address: ____________ Telephone: ____________

Manuscript number: ____________ Type of computer: ____________ Program(s) & version(s) used: ____________

Miscellaneous: ____________

**I certify that the material on the enclosed disk(s) is identical in both word and content to the printed copy herewith enclosed.**

**Signature:** ____________ **Date:** ____________

Journal of the American Society for Information Science

JASIS

## INSTRUCTIONS FOR CONTRIBUTORS

The *Journal of the American Society for Information Science (JASIS)* is a scholarly journal devoted to the various fields of documentation and information science and serves as a forum for discussion and experimentation. Manuscripts for submission to the journal (in English only) should be forwarded *in triplicate* to the Editor: Professor Donald H. Kraft, Computer Science Department, Louisiana State University, Baton Rouge, LA 70803. Books, monographs, reports, and other items for review may be sent to the Associate Editor—Book Reviews: Terrence A. Brooks, Graduate School of Library and Information Science, University of Washington, Box 352930, Seattle, WA 98195-2930. Phone: (206) 543-2646; Fax: (206) 616-3152; E-mail: tabrooks@u.washington.edu

**Types of manuscripts:** Three types of contributions are considered for publication: Full-length articles, brief communications of 1000 words or less, and letters to the editor. Letters and brief communications can generally be published sooner than full length articles. All material submitted will be acknowledged on receipt and (except for letters) subject to peer review. Copies of the referees' comments will be forwarded to the author along with the Editor's decision.

**Copyright:** No article can be published unless accompanied by a signed publication agreement, which serves as a transfer of copyright from author to publisher. A publication agreement may be obtained from the editor or the publisher. A copy of the publication agreement appears in most issues of the journal. Only original papers will be accepted and copyright in published papers will be vested in the publisher. It is the author's responsibility to obtain written permission to reproduce material that has appeared in another publication. A Permission Request Form may be obtained from the editor or the publisher.

Only articles that have not been published previously, that have not been submitted elsewhere, and that are not under review for another publication in any medium (e.g., printed journal, conference proceeding, electronic, or optical medium) should be submitted to *JASIS*. The *JASIS* Editors will assume that submission of an article to this journal implies that all the foregoing conditions are applicable.

**Format of submitted material:** All manuscripts must be submitted *in triplicate* and typed on standard letter-size bond paper (21 × 28 cm) with adequate margins. All copy, including references and captions, must be typed double spaced. The first page of the manuscript must bear the title of the paper and the full names of the authors as well as their affiliations, full addresses, telephone and fax numbers, and email addresses. In the case of multiple authors, please indicate which author is to receive correspondence and proofs. Financial support may be acknowledged in the **Acknowledgment** section at the end of the article. All succeeding pages must bear the surname of the lead author and a page number in the upper right-hand corner. An informative abstract of 200 words or less is required for articles and brief communications.

**Style:** In general, the style should follow the forms given in the *Publication Manual of the American Psychological Association* (Washington, DC: 1994). The use of metric forms is required.

**Organization:** In general, the background and purpose of the study should be stated first, followed by details of the methods, materials, procedures, and equipment used. Findings, discussion and conclusions should follow in that order. Appendices may be employed where appropriate. The APA Publication Manual should be consulted for details as needed.

**Figures:** Figures should be professionally prepared and submitted in a form suitable for reproduction (camera-ready copy). Computer-generated graphs are acceptable only if they have been printed with a good quality laser printer. All artwork must be identified on the back in light blue pencil with the figure number, author's name and address, and a shortened form of the title. Photographs should be submitted as unscreened glossy prints (20 × 25 cm). Two clearly legible copies must be supplied with each figure (these will be sent to the referees).

**Bibliography:** The accuracy and completeness of the references is the responsibility of the author. References to personal letters, papers presented at meetings, and other unpublished material may be included. If such material may be of help in the evaluation of the paper, copies should be made available to the Editor. Papers which are part of a series should include a citation of the previous paper. Explanatory material may be appended to the end of a citation to avoid footnotes in text.

The format for citations in text and for bibliographic references follows the *Publication Manual of the American Psychological Association* (4th ed., 1994). Citation of an author's work in the text should follow the author-date method of citation; the surname of the author(s) and the year of publication should appear in text. For example:

Paisley (1993) found that . . .
Recent research has shown that . . . (Schauder, 1994).
In other work (Gordon & Lenk, 1992; Harman, 1991) . . .

Examples of citations to a journal article, a book, a chapter in a book, and published proceedings of a meeting follow:

Buckland, M., & Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science, 45,* 12–19.

Borgman, C. L. (Ed.). (1990). Scholarly communication and bibliometrics. London: Sage.

Bauin, S., & Rothman, H. (1992). Impact of journals as proxies for citation counts. In P. Weingart, R. Sehringer, & M. Winterhager (Eds.), *Representations of science and technology* (pp. 225–239). Leiden: DWSO Press.

Hoppe, K., Ammersbach, K., Lutes-Schaab, B., & Zinssmeister, G. (1990). EXPRESS: An experimental interface for factual information retrieval. In J.-L. Vidick (Ed.). *Proceedings of the 13th International Conference on Research and Development in Information Retrieval (ACM SIGIR '91)* (pp. 63–81). Brussels: ACM.

In case of any question about bibliographic form, refer to the *Publication Manual of the American Psychological Association,* 4th edition, Washington, DC, 1994. Copies may be ordered from: Order Department, American Psychological Association, P.O. Box 2710, Hyattsville, MD 20784.

**Final revised manuscript:** A final version of your accepted manuscript should be submitted on diskette as well as hard copy, using the guidelines in the Diskette Submission Instructions, usually included in most issues of the journal.

**Correspondence:** All other correspondence should be addressed to the Publisher, Professional, Reference, & Trade Group, John Wiley & Sons Inc., 605 Third Ave., New York, NY 10158.

**Reprints:** Reprints of articles may be ordered from the publisher when corrected proofs are returned.

0002-8231(1998)49:13;1-I